16

# OUTLIERS, INFLUENTIAL OBSERVATIONS AND ROBUST ESTIMATION IN NON-LINEAR REGRESSION ANALYSIS AND DISCRIMINANT ANALYSIS

by

PETRUS JACOBUS UYS VAN DEVENTER



SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

IN THE DEPARTMENT OF MATHEMATICAL STATISTICS
UNIVERSITY OF CAPE TOWN
UNDER THE SUPERVISION OF PROFESSOR C.G. TROSKIE

MARCH 1993

ACKNOWLEDGEMENTS

I have to thank my supervisor, Professor Cas Troskie for providing the much needed guidance to keep me on the right track. It was his enthusiasm which always convinced me to keep at it.

I must give a special word of thanks to Trevor Hastie for making available his programs on quadratic discriminant analysis without which chapter five wouldn't have been complete.

Les Underhill supplied his SVDD-program against which I checked my own program as well as the graphical displays. It saved me a lot of time. Thank you.

I must also thank Annemé Gouws for the expert (and patient) way in which she typed this thesis.

Finally, and most importantly, I thank my wife, Danise, for her support during the last few years.

## ABSTRACT

The basic motivation for this thesis is the great difference between linear models and non-linear models in terms of the interpretability of diagnostic measures which have been developed for linear models, but have also been used frequently for the analysis of non-linear models ignoring the non-linear structure of the model.

I have found that either the linear diagnostic measures are applied directly to a data matrix as if it forms part of a linear system irrespective of the possible non-linearity of the model - see the curvature measures of Bates and Watts (1980) as well as Efron (1975) - or the model is transformed to a more or less linear form and the diagnostics derived are interpreted in terms of the parameters which have occurred after the transformation. This has the effect that the interpretation with respect to aspects of the original model may become vague.

A technique which is based on the inherent non-linear structure of the model to determine the potential presence of outliers and/or influential observations and picking them up from the data matrix is proposed. This is being achieved through the singular value decomposition of the jacobian of the model which is followed by a procedure similar to a principal components analysis.

Finally it is shown that this technique is not only useful in non-linear regression, but also in discriminant analysis where outliers and/or influential observations may cause the classification function to be not as effective as it

may be if the aberrant observations are not removed.

On the one hand I have tried to combine and clear up a variety of known facts loosely available in the literature together with proofs where none were supplied in the quoted articles, so that a concise, but understandable, readible document may be available to anyone interested in the subject of non-linear regression and diagnostics.

On the other hand I have tried to extend the knowledge available in order to show and simplify knew approaches, techniques and applications for existing (known) procedures which have not been used often and definitely not at all in non-linear regression or discriminant analysis as far as I am aware of.

---

## SOFTWARE AND HARDWARE

**SOFTWARE:** The original analysis of the non-linear model in chapter four was done by means of a set of program subroutines developed by Ratkowsky (1983). The code he supplied used routines from LINPACK which was not available at that stage so I adapted the FORTRAN-code using routines from the NUMERICAL RECIPES routines-package – see appendix D for more information on NUMERICAL RECIPES as well as the code. I found this last package very useful especially because of the well written books/manuals which were supplied with the routines. These routines have the advantage that one can edit any of them for specific personal techniques one would like to investigate.

The SVD and the computations of chapter five were done by means of a program developed specificially for the purpose of this thesis using the above-mentioned package for certain routines.

The discriminant analysis was originally done using the BMDP7M package, but as the STATGRAPHICS package is more readily available to the average man I did all further analysis using version 5.0 of this package.

**HARDWARE:** A 640K RAM, 20 MB PC was used. Some of these programs were run on an XT without a mathematical coprocessor (home) and sometimes one with a mathematical coprocessor (work).

The codes for all the self-developed programs are given in appendix D.

## CONTENTS

CHAPTER 5

INFLUENTIAL OBSERVATIONS AND OUTLIERS TOGETHER WITH COLLINEARITY IN DISCRIMINANT ANALYSIS

# CHAPTER 1

## 1. THE MODEL

The class of models under discussion is a generalization of the multiple linear regression model. The observations $\underline{y}' = (y_1,\ldots,y_n)$ are assumed to be random samples from a specified distribution $y = f(\underline{x},\underline{\theta})$ where $\underline{\theta}' = (\theta_1,\ldots,\theta_p)$ is a vector of parameters and $\underline{x}_i' = (x_1,\ldots,x_m)_i$ the ith row vector of observations of m independent variables. The observations may be raw data or functions of raw data. We assume that the distributions involved, though non-linear are continuous in nature.

The mean of the distribution of Y is $\mu = E(Y)$, i.e. the expected value is often a function of one or more of the independent variables $x_1,\ldots,x_m$. The distribution about the mean, referred to as the error distribution or random part of the model, may involve further parameters representing variances and so forth. Generally, the most commonly used distributions are the normal and gamma distributions.

Although it is hoped that observations are independent we often find some correlation among the x variables and therefore collinearity between the column vectors in the matrix of observations, $X = [x_{ij}]_{n\times m}$. By the term non-linear we mean the presence of non-linearity in the parameters of the regression function. Any model which is not of the form or can not be transformed to the form

$$Y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \ldots + \beta_p z_p + \varepsilon$$

where $\beta' = (\beta_0,\beta_1,\ldots,\beta_p)$ is the parameter vector and $z_i$ is any function of the independent (predictor) variables $x_i$, $i = 1,\ldots,m$, is called a non-linear model.

Some non-linear models however are intrinsically linear although on the surface it may look like a non-linear model. The reason for this is that it is possible to linearize the model in terms of its parameters often using a quite simple transformation - see for example the model

$$y = \exp(\theta_1 + \theta_2 x_1 + \theta_3 x_2^2 + \varepsilon)$$

which, when linearized, will have the form

$$y^* = \ell n y = \theta_1 + \theta_2 x_1 + \theta_3 x_2^2 + \varepsilon$$

A model of the form

$$y = \frac{\theta_1}{\theta_1 - \theta_2}(e^{-\theta_2 x_1} - e^{-\theta_1 x_2}) + \varepsilon$$

is intrinsically non-linear as it is impossible to linearize this model in terms of its parameters.

It is also worth to note that some authors use the words intrinsically linear only in the case where an additive error model can be linearized by means of a transformation of parameters. The model

$$y = e^{\theta x} + \varepsilon$$

can be linearized by the transformation $\beta = e^{\theta}$, i.e.

$$y^{*} = \beta x + \varepsilon$$

see eg. Draper and Smith (1981).

## 2. KEYWORDS AND PURPOSE

According to Kotz and Johnson (1983) observations are regarded as influential if their omission from the data results in substantial changes to important features of an analysis like estimates, confidence regions and test statistics. Influential observations may be outliers with large residuals relative to a specified model, or observations that are isolated from the rest of the data in the design space.

Note that an outlier is not necessarily influential, because the deletion of an outlier does not necessarily result in a change of a fitted model.

It may further be remarked that influential subsets are therefore usually those subsets which fall outside the patterns set by the majority of the data in the context of a specified model.

The reason for the observations presenting themselves as influential will not be our concern, because it will often be possible to deduce the reason of their presence as such i.e. as being influential in the context of the data and model set-up after they have been observed. Quite often however, influential data are the result of errors in data, model failure or incorrect likelihood assumptions.

Legitimate extreme observations may however also be influential and it is important to identify and report this specific type of data as well.

We shall be interested in the detection of outliers individually or as a subset of observations as well as the detection of influential observations individually or as a subset of observations. Eventually it may even be possible to detect influential variables $\underline{x}_{i'}$, where i′ consists of one or more of the i in $\underline{x}_i$, i = 1,...,m where $\underline{x}_i$ are the columns in the data matrix.

Identification of influential observations is often difficult, because jointly influential observations are not necessarily individually influential. Similarly individual influential observations are not necessarily jointly influential. To be more specific: The classical methods for the identification of outliers and influential observations do not always work in linear regression because of the masking effect due to the fact that the identification methods themselves are based on the means and covariance matrix which are also affected by these outliers and influential observations present in the data set (Rousseeuw and Van Zomeren, 1990).

During the last two decades a fair deal of effort has gone into the detection of influential observations for the general linear model $y = \underline{x}'\beta + e$, see eg. Cook and Weisberg (1982), Belsley, Kuh and Welsch (1980) as well as quite a number of other books and articles. Graphical diagnostic displays for outlying and influential observations in multiple regression are reviewed by Atkinson (1981).

## 3. STRUCTURE OF THIS THESIS

As we are dealing with the analysis of non-linear regression models it is appropriate to be able to tell whether the data we are dealing with are really non-linear in nature. It is obviously a waste of time to apply non-linear analysis techniques to linear data as the first are often much more number crunching prone then the techniques which have been derived for the analysis of linear models.

Efron (1975) showed the similarity between statistical and geometrical curvature and Bates and Watts (1980) exploited curvature further as a measure of non-linearity. In chapter two a general description of the solution to the non-linear regression problem is given using the $L_p$-norm approach as mentioned by Kennedy and Gentle (1980) and described in full by Gonin (1984), Gonin and Money (1985(i), 1985(ii)) and Gonin and Money (1989). Some suggestions are also given in terms of smoothing, penalty functions and robust estimation using new weighting factors. This is followed by a survey of Efron's description of the curvature measure with some explanatory extensions and clarifications.

In chapter three a summary is given of some of the more common measures of influence being applied in the case of the linear model and also with more or less success in the case of non-linear models. The intuitively most appealing measures here are the volume ratios measures. The drawback again is the amount of calculations necessary when applied. The one-parameter derivation of curvature of chapter two is then extended to a multiple parameter derivation by means of a differential geometrical approach (Ross, 1984). Bates and Watts' description (1980) gives further insight into so-called relative curvature

measures by means of which inferences can be made regarding the appropriateness of non-linear versus linear diagnostic techniques.

This chapter then is completed by a summary of other diagnostics than the direct approach, i.e. the application of linear diagnostics on a "linear" model when the chances are good that the model is non-linear in nature. In this section the fact that the data matrix in the linear model is in fact the first derivative of the model $y = \beta' \underline{x} + \varepsilon$ is applied. That means that the jacobian of the non-linear model is used in some (or all ?) of the linearly derived techniques.

In chapter four a new approach to diagnostics is derived which is more appropriate for use in the non-linear situation. The jacobian of the model is determined and an analysis similar to a principal components analysis giving ratios of variance contributions is applied based on the jacobian. This approach creates the possibility of determining outliers, influential observations as well as collinear variables. This chapter is concluded by an example in which the application of the above mentioned is demonstrated. Mention is also being made of the Mallows robust technique and its applicability in dealing with diagnostics.

Chapter five reviews some existing techniques with adaptations for diagnostics in discriminant analysis and then takes some of the suggestions of the previous chapter further by suggesting the application of the above-mentioned technique to a discriminant analysis in order to determine outliers and influential observations which may influence the discriminatory power of the resulting classification function(s) badly.

At the end of this chapter two examples are worked extensively. These examples show the advantages of the new approach, i.e. picking up possible outliers/influential observations, irrespective of how many of the latter there may be (within realistic limits). We also compare the results of the linear discriminant approach of Fisher to other more recent techniques.

This chapter is concluded by suggestions for further research into the use(s) of the singular value decomposition (SVD) technique as well as a recap of other suggestions which have occurred in the thesis.

---

# CHAPTER 2

## A SURVEY OF NON-LINEAR REGRESSION AND CURVATURE

### 2.1 THE NON-LINEAR REGRESSION MODEL AND SOME SOLUTIONS

In recent years more interest has been shown in the solution of the non-linear regression problem and its associated diagnostics. One of the main reasons being the availability as well as number-crunching ability of main-frame computers and even the modern personal computers.

The non-linear regression model can be defined as usual: Let

$$y_i = f(\underline{x}_i, \underline{\theta}) + \varepsilon_i \qquad 2.1.1$$

where $y_i$ is the ith observation of $\underline{Y}$.

$\underline{x}_i'$ is the ith observation vector = $(x_{1i},\ldots,x_{mi})$

$\varepsilon_i$ is the ith unknown error $i = 1,\ldots,n$

$\underline{\theta}' = (\theta_1,\ldots,\theta_k)$, $n > k$ in general.

When $\varepsilon_i \sim n(0,\sigma^2)$ the $L_2$-method of solution is applicable. If, however $\varepsilon_i$ is not distributed normal, $L_2$ is not necessarily applicable. In such a case the $L_p$-norm, $p\neq 2$ may be used where p has some relationship with the coefficient of kurtosis (Gonin and Money, 1985(i) and (ii)). The $L_p$-norm is used in the following way: Determine $\underline{\theta}$ in such a way that $S_p(\underline{\theta})$ is being minimized:

$$S_p(\underline{\theta}) = \sum_{i=1}^{n} | y_i - f(\underline{x}_i, \underline{\theta})|^p \qquad 2.1.2$$

A combination of methods can be applied to determine the minimum of $S_p(\underline{\theta})$, e.g.:

If $S_p(\underline{\theta}) = \Sigma F_i$

i.e. $F_i = |Y_i - f(\underline{X}_i, \underline{\theta})|^p$ 2.1.3

then it can be shown that the minimum may be obtained by the following, see Gonin (1984) and Gonin and Money (1989):

The p-Jacobian matrix is

$$J_p = \left[|y_i - f(\underline{x}_i,\underline{\theta})|^{\frac{1}{2}p-1} \frac{\partial f_i}{\partial \theta_j}\right]_{n\times k}, \quad \begin{matrix} i=1,\ldots n \\ j=1,\ldots,k \end{matrix} \qquad 2.1.4$$

The "residual vector" is written as

$$\begin{aligned}(\underline{y} - \underline{f}(\underline{x},\underline{\theta})) &= (\underline{y} - \underline{f})_p \\ &= \left[|y_i - f_i|^{\frac{1}{2}p-1} (y_i - f_i)\right]_{n\times 1}, \quad i=1,\ldots,n \end{aligned} \qquad 2.1.5$$

In the following care must be taken in order to distinguish between differentiation of $f_i$ on the one hand and $S_p(\underline{\theta})$ on the other. The aim is to write the derivatives of $S_p(\underline{\theta})$ in terms of the derivatives of $f_i$.

If we define

$$B_p(\underline{\theta}) = \sum_{j=1}^{n} |y_i - f_i|^{p-2} (f_i - y_i)\nabla^2 f_i \ , \qquad 2.1.6$$

where $\nabla^2 f_i$ is the Hessian of $f_i$, i.e.

$$\nabla^2 f_i = \begin{bmatrix} \frac{\partial^2 f_i}{\partial \theta_1^2} & \frac{\partial^2 f_i}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 f_i}{\partial \theta_1 \partial \theta_k} \\ \cdots\cdots\cdots & & & \cdots\cdots \\ \frac{\partial^2 f_i}{\partial \theta_k \partial \theta_1} & & \cdots\cdots\cdots & \frac{\partial^2 f_i}{\partial \theta_k^2} \end{bmatrix}_{k \times k} \qquad 2.1.7$$

then the first - and second order partial derivatives of $S_p(\underline{\theta})$ with respect to $\underline{\theta}$ are:

$$\nabla S_p(\underline{\theta}) = -pJ_p'(\underline{y} - \underline{f})_p \qquad 2.1.8$$

$$\nabla^2 S_p(\underline{\theta}) = p\left[(p-1)J_p'J_p + B_p(\underline{\theta})\right] \qquad 2.1.9$$

with

$$\frac{\partial S_p(\underline{\theta})}{\partial \theta_\ell} = \sum_{i=1}^{n} -p|y_i - f_i|^{p-1}(y_i - f_i)\frac{\partial f_i}{\partial \theta_i}, \qquad \ell = 1,\ldots,k \qquad 2.1.10$$

$$\frac{\partial^2 S_p(\underline{\theta})}{\partial \theta_i \partial \theta_s} = p\sum_{i=1}^{n} |y_i - f_i|^{p-2}\left\{(p-1)\frac{\partial f_i}{\partial \theta_\ell}\frac{\partial f_i}{\partial \theta_s} + (y_i - f_i)\frac{\theta^2 f_i}{\partial \theta_\ell \partial \theta_s}\right\} \qquad 2.1.11$$

The above mentioned can be proved as follows:

$$F_i = |y_i - f(\underline{x}_i, \underline{\theta})|^p$$

$$\frac{\partial F_i}{\partial \theta_j} = -p \operatorname{sign}(y_i - f(\underline{x}_i, \underline{\theta}))\ |y_i - f(\underline{x}_i, \underline{\theta})|^{p-1} \frac{\partial f_i}{\partial \theta_j}$$

$$= -p \frac{y_i - f(\underline{x}_i, \underline{\theta})}{|y_i - f(\underline{x}_i, \underline{\theta})|}\ |y_i - f(\underline{x}_i, \underline{\theta})|^{p-1} \frac{\partial f_i}{\partial \theta_j}$$

$$= -p(y_i - f(\underline{x}_i, \underline{\theta}))\ |y_i - f(\underline{x}_i, \underline{\theta})|^{p-2} \frac{\partial f_i}{\partial \theta_j} \qquad 2.1.12$$

$$\frac{\partial^2 F_i}{\partial \theta_i \partial \theta_s} = -p|y_i - f_i|^{p-2}(y_i - f_i) \frac{\partial^2 f_i}{\partial \theta_\ell \partial \theta_s} + p|y_i - f_i|^{p-2} \frac{\partial f_i}{\partial \theta_\ell} \frac{\partial f_i}{\partial \theta_s}$$

$$+ p(p-2)|y_i - f_i|^{p-4} (y_i - f_i)^2 \frac{\partial f_i}{\partial \theta_\ell} \frac{\partial f_i}{\partial \theta_s}$$

$$= p(p-1)|y_i - f_i|^{p-2} \frac{\partial f_i}{\partial \theta_\ell} \frac{\partial f_i}{\partial \theta_s} - p|y_i - f_i|^{p-2} (y_i - f_i) \frac{\partial^2 f_i}{\partial \theta_\ell \partial \theta_s} \qquad 2.1.13$$

Note again that the $L_2$-norm is imbedded in the $L_p$-norm. Let $p = 2$ then

$$\nabla S_2(\underline{\theta}) = -2J'(\underline{y} - \underline{f}) \qquad 2.1.14$$

$$\nabla^2 S_2(\underline{\theta}) = 2(J'J + B_2(\underline{\theta})). \qquad 2.1.15$$

It can further be seen that

$$(\underline{f} - \underline{y})_p = \left[|f_i - y_i|^{\frac{1}{2}p-1} (f_i - y_i)\right] \qquad i=1,\ldots,n$$

$$= -\left[|f_i - y_i|^{\frac{1}{2}p-1}(y_i - f_i)\right]$$

$$= -(\underline{y} - \underline{f})_p \qquad 2.1.16$$

with the result that

$$\nabla S_p(\underline{\theta}) = -pJ_p'(\underline{y} - \underline{f})_p$$

$$= pJ_p'(\underline{f} - \underline{y})_p \qquad 2.1.17$$

Note that if p = 1 with one of the parameters in the role of a constant, e.g. $f = \theta_0 + \theta_2 e^{\theta_1 x}$ then the second order methods are not applicable, because $\nabla^2 S_1(\underline{\theta})$ will become singular. In the case of p = 1 the method of Osborne and Watson (1971) may be used. A second order Taylor series expansion of $S_p(\underline{\theta})$ about $\underline{\theta}^j$, where $\underline{\theta}^j$ is the jth estimate, now yields:

$$S_p(\underline{\theta}) \approx S_p(\underline{\theta}^j) + (\underline{\theta} - \underline{\theta}^j)'\nabla^2 S_p(\underline{\theta}^j) + \tfrac{1}{2}(\underline{\theta} - \underline{\theta}^j)'\nabla^2 S_p(\underline{\theta}^j)(\underline{\theta} - \underline{\theta}^j) + \ldots \qquad 2.1.18$$

A necessary condition for $\underline{\theta}$ to be a local minimum of $S_p(\underline{\theta})$ is:

$$\nabla S_p(\underline{\theta}) = 0 \qquad 2.1.19$$

i.e.

$$\nabla S_p(\underline{\theta}) = 0 = 0 + \nabla S_p(\underline{\theta}^j) + (\underline{\theta} - \underline{\theta}^j)'\nabla^2 S_p(\underline{\theta}^j)$$

$$\nabla S_p(\underline{\theta}^j) + \nabla^2 S_p(\underline{\theta}^j)(\underline{\theta} - \underline{\theta}^j) \approx 0 \qquad 2.1.20$$

i.e.

$$\underline{\theta} - \underline{\theta}^j = \underline{d}^j = -\left[\nabla^2 S_p(\underline{\theta}^j)\right] \nabla S_p(\underline{\theta}^j) \qquad 2.1.21$$

or

$$\underline{d}^j = -p\left[p(p-1)J_p' J_p + pB_p(\underline{\theta})\right]^{-1} J_p'(\underline{f} - \underline{y})_p$$

$$= -\left[(p-1)J_p' J_p + B_p(\underline{\theta})\right]^{-1} J_p'(\underline{f} - \underline{y})_p \qquad 2.1.22$$

i.e.

$$\underline{\theta}^{j+1} = \underline{\theta}^j + \left[(p-1)J_p' J_p + B_p(\underline{\theta})\right]^{-1} J_p'(\underline{f} - \underline{y})_p \qquad 2.1.23$$

as $(\underline{y} - \underline{f})_p = -(\underline{f} - \underline{y})_p$.

The numerical solution using the Choleski factorization and singular value decomposition for certain conditions does not concern us at the present. Suffice it to say that these techniques are used to determine $\underline{d} = \underline{d}_1 + \underline{d}_2$ where $\underline{d}_1$ is a vector of values obtained in a straightforward way when there is no complications with the eigen roots of $J_p' J_p$ and $\underline{d}_2$ the values used corresponding to the indefinite or negative eigen values of $J_p'$.

It is interesting to note that for $p = 2$ we have again the least squares solution, but now with information from the second order derivative as well.

In this form the task to program for a solution can be formidable because of the hessian in $B_p(\underline{\theta})$. Besides this it can be very difficult to find the required second derivatives either analytically or by means of finite diffences. Furthermore, the presence of $B_p(\underline{\theta})$ does not always lead to quicker convengence. When the residuals are large however, or in the case of fairly ill-conditioned problems $B_p(\underline{\theta})$ can not be ignored. (Gill and Murray, 1978 and Nazareth, 1980).

Marquardt used an approximation for $B_2(\underline{\theta})$, i.e. $\lambda I$ which leads to a compromise between the methods of steepest descent (where $\lambda \rightarrow \infty$) and the Gauss-Newton approach ($\lambda \rightarrow 0$). This corresponds to an angle of descent of between 90° ($\lambda \rightarrow \infty$) and 0°($\lambda \rightarrow 0$) relative to the tangent to a contour created by a specific constant $S(\underline{\theta})$ (Marquardt, 1963).

If the $L_p$ approach is being used the question may be asked about which p-value is most appropriate. Gonin and Money (1985), Gonin (1984(ii)), Sposito (1982) and Money et al.(1982) showed that there is a relationship between p and the coefficient of kurtosis; i.e.

$$p = \frac{9}{\beta_2^2} + 1$$

and

$$p = \frac{6}{\beta^2}$$

depending on sample size, so that an adaptive scheme may be used. This will be so if the asymptotic distribution of p with mean equal to 2 is ignored. The asymptotic characteristic is showing up at $n \geq 30$ already and becomes very strong at $n \geq 200$.

In the case that the derivatives are very difficult to determine the so-called Quasi-Newton techniques may be applied where the derivatives are determined by means of numerical difference techniques.

Good examples of a practical approach to the solution of this type of problem can be found in Kennedy and Gentle (1980) as well as Ratkowsky (1983). In the first some mention was made of the use of the $L_p$-model, $p \neq 2$, which was taken further theoretically by Gonin (1984). Ratkowsky went further in analyzing some practical problems especially with regard to intrinsic and parametric curvature and the calculation of these two components in a non-linear model.

At this point it is of interest to refer to Hastie and Tibshirani (1990). They show how the cubic smoothing splines emerge as the solution to an optimization problem.

Given all functions f(x) with continuous first and second derivatives one must find the function which minimizes the penalized residual sum of squares

$$\sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int_a^b \left\{f''(t)\right\}^2 dt \qquad 2.1.24$$

where $\lambda$ is a fixed constant and x is ordered such that $a \leq x_1 \leq x_2 \leq \ldots \leq x_n \leq b$. This satisfies the scatterplot smoother requirements as mentioned by Hastie et al. The second term in 2.1.24 is known as the penalty function. This way of representing the problem of optimization introduces us to the subject of linear and non-linear equality or inequality constraints which can be given in the form

$$\underline{\theta}'\underline{\theta} \leq d^2$$

d being a constant, or in more general terms

$$\underline{\theta}'\ \Omega\ \underline{\theta} \leq d^2\ . \qquad 2.1.25$$

In the argument of Hastie et al.:-

If the kernel smooth at $x_0$ is

$$S(x_0) = \frac{\sum_{i=1}^{n} d(\frac{x_0 - x_i}{x})y_i}{\sum_{i=1}^{n} d(\frac{x_0 - x_i}{x})}, \qquad 2.1.26$$

and $\underline{S}(x)$ is written in the vector form as

$$\underline{S}(x) = \sum_{j=1}^{n} \gamma_j B_j(x) \qquad 2.1.27$$

with $B_j$ the cubic B-spline basis and $\underline{S}$ in a 2-dimensional space, and further if $\underline{f}$ is replaced by $\underline{S}$ then the $(n+2)\times(n+2)$ matrix $\Omega$ is

$$\Omega = [B_{ij}] = [B_j(x_i)]$$

$$= \int B_i''(x)\ B_j''(x)\ dx \qquad 2.1.28$$

so that 2.1.24 becomes

$$(\underline{y} - B\underline{\gamma})'(\underline{y} - B\underline{\gamma}) + \lambda\underline{\gamma}'\Omega\underline{\gamma} \qquad 2.1.29$$

The boundary derivative constraints are now automatically imposed by the penalty term.

For an optimum solution the derivative with respect to $\underline{\gamma}$ must be equal to zero so that

$$(B'B + \lambda\Omega)\hat{\underline{\gamma}} = B'\underline{y}$$

i.e.

$$\hat{\underline{\gamma}} = (B'B + \lambda\Omega)^{-1}B'\underline{y} \qquad 2.1.30$$

If N is a non-singular natural-spline basis matrix for the solution of 2.1.27 and $\hat{\underline{\beta}}$ the transformed version of $\hat{\underline{\gamma}}$ which corresponds to a change in basis, we find that

$$\underline{S} = N\hat{\underline{\beta}}$$

$$= N(N'N + \lambda\Omega)^{-1}N'\underline{y}$$

$$= (I + \lambda K)^{-1}\underline{y} \qquad 2.1.31$$

with $K = N'\Omega N^{-1}$. This means that the cubic smoothing spline $\underline{S}$ minimizes

$$(\underline{y} - \underline{f})'(\underline{y} - \underline{f}) + \lambda\underline{f}'K\underline{f} \qquad 2.1.32$$

where $\underline{f}'K\underline{f}$ is called a roughness penalty which is a quadratic form in second differences.

What is interesting here, is the obvious correspondence between the form of the equation 2.1.22 and 2.1.30. Although Hastie et al. showed this for a single independent variable, it should be easy to extend it to a multi-independent variable situation. The only problem would be to find an ordering convention as indicated with 2.1.24.

At the moment this aspect of constraints and the penalty function is due for further research by the author.

We have besides these methods of solving the non-linear regression problem a number of recent, and not so recent developments. One of these is the Mallows-Type bounded-influence-regression approach as proposed by De Jongh, De Wet and Welsh (1988). There approach stems from the regression trimmed mean estimators of Koenker and Basset (1978) and Mallows (1973 and 1975). De Jongh et al however, include a trim on the independent variables as well as an extension on the earlier approaches. In chapter four and five more detail is given together with some empirical work and comparisons and further remarks.

Friedman (1991) gives an extended summary of (robust) estimation in (non)-linear regression. Weighting plays an important role in some of these techniques. In this regard I want to propose some weights which I do not follow up here, but which are being investigated:

When considering weights for the minimization problem of particular interest may

be to use the studentized residuals or $DFFITS_i$ instead of least squares residuals (i.e. $L_p$-norm).

$$DFFITS_i: \quad \Sigma \frac{|e_i|^2}{s^2_{(i)}(1-h_i)} \cdot \frac{h_i}{1-h_i}$$

$$= \Sigma \frac{h_i}{s^2_{(i)}(1-h_i)^2} |e_i|^2$$

$$\text{or:} \quad \Sigma \left\{ \frac{h_i^{\frac{1}{2}}}{s_{(i)}(1-h_i)} \right\}^p |e_i|^p \qquad 2.1.33$$

with $s_{(i)}$ the "leave out" standard deviation. The ith diagonal element of $X(X'X)^{-1}X'$ or the matrix $J(J'J)^{-1}J'$ will be represented by $h_i$ or $j_i$ respectively, depending on whether the model is linear or non-linear.

Similarly for the studentized residual:-

$$\Sigma \left| \frac{e_i}{s\sqrt{1-h_i}} \right|^p$$

$$\text{or} \qquad \Sigma \left| \frac{e_{(i)}}{s_{(i)}\sqrt{1-h_i}} \right|^p, \text{ p probably = 2,} \qquad 2.1.34$$

with $e_{(i)}$ being the predicted residual, i.e. $y_i - \underline{x}_i'\hat{\underline{\beta}}_{(i)}$. The respective weights are thus

$$\frac{h_i}{s^2_{(i)}(1-h_i)^2} \text{ or } \left[ \frac{h_i^{\frac{1}{2}}}{s_{(i)}(1-h_i)} \right]^p$$

and

$$(s\sqrt{1 - h_i})^{-p} \text{ or } (s_{(i)}\sqrt{1 - h_i})^{-p} \qquad 2.1.35$$

where $h_i$ may be replaced by $j_i$ in 2.1.34 as well as 2.1.35.

The weighting however does not necessary have to be applied as an absolute weight, but can be applied in the sense of a regularization type of constant weighting the suspect observation - see Friedman (1989). A regularized ridge based on the second term of the inverted matrix in the solution of $\underline{\theta}$ in 2.1.23 is applied, i.e.

$$\underline{\theta}^{j+1} = \underline{\theta}^{j} + \left[(p-1)J'_{p(i)}J_{p(i)} + kj_i j_i'\right]^{-1} J'_{p(i)} (\underline{f} - \underline{y})_p \qquad 2.1.36$$

where $0 \leq k \leq 1$. Obviously, when $k = 1$ the solution is almost similar to the Gauss-Newton solution. When $k < 1$ the structure of the (non)-linear model is retained (in a lesser way) but the heavy load of calculations involved in $B_p(\underline{\theta})$ falls away. This form of the solution and the results are currently being investigated.

When the subject of diagnostics comes up the masking effect plays an important role. To counter this Rousseeuw and Van Zomeren (1990) propose the application of estimators of multivariate location and covariance values that have high

break down points. Considering the recognition of bad or good high leverage points they show that a relative large $h_i$ which is the ith diagonal element of $X(X'X)^{-1}X'$ does not necessarily indicate high leverage. They propose the plot of the robust Mahalanobis distance (i.e. ignoring the observations under the breakdown point) against the standardized least median of squares (LMS). From this one should be able to distinguish between vertical, good and bad leverage observations.

Finding the breakdown point, i.e. the fraction of observations (outliers) which may be left out of the analysis before the analysis as such breaks down, is discussed in detail in Lopuhaä and Rousseeuw (1991).

## 2.2 CURVATURE AND NON-LINEAR REGRESSION

### 2.2.1 Statistical curvature

The statistical curvature of a one-parameter family of probability density functions $F = \{f(\underline{x},\theta)\}$, can be defined in the following way (Efron, 1975):

The statistical curvature of F at $\theta$ measures how near F is to being exponential and is indicated by the symbol $\gamma_\theta$. If $\gamma_\theta$ is equal to zero F is exponential. Otherwise $\gamma_\theta > 0$ for at least some values of $\theta$. It can be shown that $\gamma_\theta^2$ gives an indication of the amount of information lost when maximum likelihood estimation (m.l.e.) is used in an exponential situation. Refer Cox and Hinkley (1974) on the information number.

Note that

$$I(\theta) = E\left(\frac{\partial}{\partial\theta} \log f(\underline{x},\theta)\right)^2$$

$$= \int_{-\infty}^{\infty}\int_{\infty}^{\infty} \left(\frac{\partial}{\partial\theta} \log f(\underline{x},\theta)\right)^2 f(\underline{x},\theta)\, d\theta$$

$$= \int_{-\infty}^{\infty}\int_{\infty}^{\infty} \frac{1}{f(\underline{x},\theta)} \left(\frac{\partial}{\partial\theta} f(\underline{x},\theta)\right)^2 d\theta$$

$$= \mathrm{var}\left(\frac{\partial}{\partial\theta} f(\underline{x},\theta)\right) \qquad 2.2.1.1$$

Following Fisher (1925) and Rao (1963): If F is a one parameter subset of the k-category multinomial distributions, indexed by the vector of probabilities $f(\underline{x},\theta) = P_\theta$ (X $\in$ category x), x = 1,...,k, then

$$\gamma^2_\theta = \lim_{n\to\infty} \left(\underline{i}_\theta - \underline{i}^{\hat{\theta}}_\theta\right)$$

$$= \underline{i}_\theta \left\{ \frac{\mu_{02} - 2\mu_{21} + \mu_{40}}{i^2_\theta} - 1 - \frac{\mu^2_{11} + \mu^2_{30} - 2\mu_{11}\mu_{30}}{i^3_\theta} \right\} \qquad 2.2.1.2$$

is the asymptotic loss of information, where $\underline{i}_\theta$ is the Fisher information in an independent sample of size n from $f(\underline{x},\theta)$,

$\underline{i}^{\hat{\theta}}_\theta$ is the Fisher information in the m.l.e., i.e. in $\hat{\theta}(\underline{x}_1,\ldots,\underline{x}_n)$,

i is the Fisher information in a sample of size one, i.e. $\underline{i}_\theta = ni_\theta$, and

$$\mu_{hj} = E_\theta\left\{\frac{\dot{f}(\underline{x},\theta)}{f(\underline{x},\theta)}\right\}^h \left\{\frac{\ddot{f}(\underline{x},\theta)}{f(\underline{x},\theta)}\right\}^j, \quad \text{where } \dot{f} = \frac{\partial f}{\partial\theta} \text{ and } \ddot{f} = \frac{\theta^2 f}{\partial\theta^2}\ .$$

For any other consistent efficient estimator $T(x_1,\ldots,x_n)$ the asymptotic loss of information, i.e.

$$\lim_{n\to\infty}(\underline{i} - \underline{i}^T_\theta) \geq \lim_{n\to\infty}(\underline{i} - \underline{i}^{\hat{\theta}}_\theta), \qquad 2.2.1.3$$

which is called the "second order efficiency property" of the m.l.e..

Efron stated further that if F is the set of families which are subsets of multi-parameter exponential families then if the subset forms a straight line in the natural parameter space of F, F is a one-parameter exponential family. A curved line subset indicates that F is not exponential and in such a case the statistical curvature equals the ordinary geometrical curvature of the line exactly, where the latter shows the rate of charge of direction with respect to arc-length.



## 2.2.2 Geometrical curvature

For a definition of curvature in the Euclidian k-space, $E_k$, define the curved line L as follows

$$L = \{\eta_\theta,\ \theta \in \Theta\} \qquad 2.2.2.1$$

where $\Theta$ is an interval of the real line, i.e. $\theta$ is one of an infinite number of

elements in $\theta$.

For each $\theta$, $\eta_\theta$ is a vector in $E^k$, i.e. $\eta_\theta = \eta(\underline{\theta})$ forms the components (points) of L in space. Further, the derivatives

$$\dot{\eta}_\theta = \frac{\partial \eta_\theta}{\partial \theta}, \quad \ddot{\eta}_\theta = \frac{\partial^2 \eta_\theta}{\partial \theta^2} \qquad 2.2.2.2$$

exist continuously in a neighbourhood of a value of $\theta$ where the curvature will be defined.

Consider now the semi positive definite matrix $\Sigma_\theta$:k×k, which is defined continuously in $\theta$, i.e. each component of $\Sigma_\theta$ is a continuous function of $\theta$. Define the 2×2-matrix $\mathbf{M}_\theta$ as follows:

$$\mathbf{M}_\theta = \begin{bmatrix} \nu_{20}(\theta) & \nu_{11}(\theta) \\ \nu_{11}(\theta) & \nu_{02}(\theta) \end{bmatrix}$$

$$= \begin{bmatrix} \dot{\eta}'_\theta \ \Sigma_\theta \ \dot{\eta}_\theta & \dot{\eta}'_\theta \ \Sigma_\theta \ \ddot{\eta}_\theta \\ \ddot{\eta}'_\theta \ \Sigma_\theta \ \dot{\eta}_\theta & \ddot{\eta}'_\theta \ \Sigma_\theta \ \ddot{\eta}_\theta \end{bmatrix} \qquad 2.2.2.3$$

then

$$\gamma_\theta = (|\mathbf{M}_\theta| / \nu_{20}^3(\theta))^{1/2} \qquad 2.2.2.4$$

is the curvature of L at $\theta$ with respect to the inner product of $\Sigma_\theta$.

Thus $\gamma_\theta$ is the rate of change of direction of $\eta_\theta$ with respect to arc-length

along L.

### 2.2.3 The correspondence between statistical and geometrical curvature

To find the correspondence between statistical and geometrical curvature let us have a look at the statistical curvature of one parameter families F which are curved subsets of a larger k-parameter exponential family by introducing

$$g_{\underline{\eta}}(\underline{x}) \equiv g(\underline{x})e^{\underline{\eta}'\underline{x} - \varphi(\underline{\eta})} \qquad 2.2.3.1$$

which is a family of densities with respect to a measure m(.) on the Euclidian k-space $E^k$, $\varphi(\underline{\eta})$ shortly to be defined, and

$$\underline{\eta} \in H \subset E^k \ni \int_{E^k} g(\underline{x})e^{\underline{\eta}'\underline{x}}dm(\underline{x}) < \infty \qquad 2.2.3.2$$

where H is convex and is called the "natural parameter space" of the exponential family. Define further

$$\left[\frac{\partial}{\partial\eta_i}\varphi(\underline{\eta})\right] = \underline{\lambda}(\underline{\eta}) \equiv E_{\underline{\eta}}(\underline{x}) \qquad 2.2.3.3$$

as can be seen in the following way. Let

$$\int_{E^k} g(\underline{x})e^{\underline{\eta}'\underline{x} - \varphi(\underline{\eta})}dm(\underline{x}) = 1,$$

then

$$\int_{E^k} g(\underline{x})e^{\underline{\eta}'\underline{x}}dm(\underline{x}) = e^{\varphi(\underline{\eta})}$$

$$\frac{\partial}{\partial\eta_i}\left(e^{\varphi(\underline{\eta})}\right) = \frac{\partial}{\partial\eta_i}\int_{E^k} g(\underline{x})e^{\underline{\eta}'\underline{x}}dm(\underline{x})$$

$$\frac{\partial}{\partial\eta_i}\left(\varphi(\underline{\eta})\right)e^{\varphi(\underline{\eta})} = \int_{E^k} g(\underline{x})x_i e^{\underline{\eta}'\underline{x}}dm(\underline{x})$$

$$\begin{aligned}\frac{\partial}{\partial\eta_i}\left(\varphi(\underline{\eta})\right) &= \int_{E^k} x_i g(\underline{x})e^{\underline{\eta}'\underline{x} - \varphi(\underline{\eta})}dm(\underline{x}) \\ &= E(x_i) \\ &= \lambda_i(\underline{\eta}) \qquad \text{2.2.3.4}\end{aligned}$$

so that $\lambda_i(\underline{\eta}) = E(x_i) = \frac{\partial}{\partial\eta_i}\varphi(\underline{\eta})$, or

$$\underline{\lambda}(\underline{\eta}) = E_{\underline{\eta}}(\underline{x}) = \left[\frac{\partial}{\partial\eta_i}\varphi(\underline{\eta})\right] \qquad \text{2.2.3.5}$$

and

$$\frac{\partial^2\varphi(\underline{\eta})}{\partial\eta_j\partial\eta_i} = \int_{E^k}\left(x_j - \frac{\partial\varphi(\underline{\eta})}{\partial\eta_j}\right) x_i\, g(\underline{x})e^{\underline{\eta}'\underline{x} - \varphi(\underline{\eta})}dm(\underline{x})$$

$$= \int_{E^k} x_j x_i g(\underline{x})e^{\underline{\eta}'\underline{x} - \varphi(\underline{\eta})}dm(\underline{x}) - \frac{\partial\varphi(\underline{\eta})}{\partial\eta_j}\int_{E^k} x_i\, g(\underline{x})e^{\underline{\eta}'\underline{x} - \varphi(\underline{\eta})}dm(\underline{x})$$

$$= E(x_j x_i) - E(x_j)E(x_i)$$

$$= E[x_j - E(x_j)]E[x_i - E(x_i)] \text{ i.e.}$$

$$\Sigma = \mathrm{cov}_{g_\eta}(\underline{x}) = \frac{\partial^2 \varphi(\eta)}{\partial \eta_i \partial \eta_j} \qquad 2.2.3.6$$

Let us indicate

$$\Lambda = \{\underline{\lambda}(\eta): \eta \in H\} \qquad 2.2.3.7$$

which is a one-to-one mapping from H to $\Lambda$ so that $\lambda = \lambda(\eta)$.

If $\Sigma$ has rank $r \geq 2$ to prevent any trivialities, let

$$L = \{\eta_\theta: \theta \in \Theta\}$$

be a one parameter set in H, with $\eta_\theta$ a twice differentiable function of $\theta$ and

$$f_\theta(\underline{x}) \equiv g_\eta(\underline{x}) = g(\underline{x}) e^{\eta' \underline{x} - \varphi_\theta} \qquad 2.2.3.8$$

where $\varphi_\theta = \varphi(\eta_\theta)$, $\underline{\lambda}_\theta = \underline{\lambda}(\eta_\theta)$ and $\Sigma_\theta = \Sigma(\eta_\theta)$. One can show that

$$\Sigma_\theta \dot{\eta}_\theta = \underline{\lambda}_\theta \text{ and } \dot{\varphi}_\theta = \dot{\eta}_\theta' \underline{\lambda}_\theta = E_\theta[\dot{\eta}_\theta' \underline{x}] \ . \qquad 2.2.3.9$$

F will be used for the curved exponential family of densities $\{f_\theta(\underline{x}): \theta \in \Theta\}$.

We can define now $\gamma_\theta$ as the statistical curvature of F at $\theta$ which is just the geometrical curvature of $L = \{\eta_\theta: \theta \in \Theta\}$ at $\theta$ with respect to the covariance inner product $\Sigma_\theta$.

If L is a straight line through H, say $\eta = a + b\tau(\theta)$ with a and b known, but $\tau(\theta)$ is a real-valued twice differentiable function of $\theta$, then $\gamma_\theta = 0\ \forall\theta$, because the curvature of a straight line is zero. From this follows that all members of

$$f_\theta(\underline{x}) = g(\underline{x})e^{\underline{a}'\underline{x}}\, e^{\tau(\theta)\underline{b}'\underline{x}} - \varphi_\theta \qquad 2.2.3.10$$

which is a one-parameter exponential family with natural parameter $\tau(\theta)$ and sufficient statistic $\underline{b}'\underline{x}$, have statistical curvature everywhere equal to zero.

Let $\ell_\theta(\underline{x}) \equiv \log f_\theta(\underline{x})$

so that

$$\dot{\ell}_\theta(\underline{x}) \equiv \frac{\partial}{\partial\theta}\ell_\theta(\underline{x}),\ \ddot{\ell}_\theta(\underline{x}) \equiv \frac{\partial^2}{\partial\theta^2}\ell_\theta(\underline{x}) \qquad 2.2.3.11$$

$$E_\theta[\dot{\ell}_\theta(\underline{x})] = E\left[\frac{\partial}{\partial\theta}\left(\log g(\underline{x}) + \underline{\eta}'_\theta\underline{x} - \varphi(\underline{\eta}_\theta)\right)\right]$$

$$= E[(\dot{\underline{\eta}}'_\theta\underline{x} - \dot{\underline{\eta}}'_\theta\underline{\lambda}_\theta)]$$

$$= \dot{\underline{\eta}}'_\theta\underline{\lambda}_\theta - \dot{\underline{\eta}}'_\theta\underline{\lambda}_\theta = 0 \qquad 2.2.3.12$$

and

$$E_\theta\left[\ddot{\ell}_\theta(\underline{x})\right] = E\left[\frac{\partial}{\partial\theta}\left(\dot{\underline{\eta}}'_\theta\underline{x} - \dot{\underline{\eta}}'_\theta\underline{\lambda}_\theta\right)\right]$$

$$= E\left[\frac{\partial}{\partial\theta}\,\dot{\eta}'_\theta\,(\underline{x} - \underline{\lambda}_\theta)\right]$$

$$= E\left[\ddot{\eta}'_\theta\,(\underline{x} - \underline{\lambda}_\theta) - \dot{\eta}'_\theta\Sigma_\theta\dot{\eta}_\theta\right]$$

$$= 0 - E\left[\dot{\eta}'_\theta\Sigma_\theta\dot{\eta}_\theta\right]$$

$$= -E_\theta[\ell_\theta^2] \equiv -i_\theta \qquad 2.2.3.13$$

or rather

$$E_\theta[\ell_\theta^2] = -E_\theta[\ddot{\ell}_\theta] = i_\theta \qquad 2.2.3.14$$

where $i_\theta$ is Fisher's information. Now as $i_\theta = \dot{\eta}'_\theta\Sigma_\theta\dot{\eta}_\theta$, the matrixcov$(\dot{\ell}_\theta, \ddot{\ell}_\theta)$ can be expressed as - keeping in mind that $\ddot{\ell}_\theta + i_\theta = \ddot{\eta}'_\theta(\underline{x} - \underline{\lambda}_\theta)$:

$$\text{covmatrix}(\dot{\ell}_\theta, \ddot{\ell}_\theta) = E\begin{bmatrix}\dot{\ell}_\theta - E(\dot{\ell}_\theta) \\ \ddot{\ell}_\theta - E(\ddot{\ell}_\theta)\end{bmatrix}\begin{bmatrix}\dot{\ell}_\theta - E(\dot{\ell}_\theta) & \ddot{\ell}_\theta - E(\ddot{\ell}_\theta)\end{bmatrix}$$

$$= E\begin{bmatrix}\dot{\ell}_\theta \\ \ddot{\ell}_\theta + i_\theta\end{bmatrix}\begin{bmatrix}\dot{\ell}_\theta & \ddot{\ell}_\theta + i_\theta\end{bmatrix}$$

$$= E\begin{bmatrix}\dot{\eta}'_\theta(\underline{x} - \underline{\lambda}_\theta) \\ \ddot{\eta}'_\theta(\underline{x} - \underline{\lambda}_\theta)\end{bmatrix}\begin{bmatrix}\dot{\eta}'_\theta(\underline{x} - \underline{\lambda}_\theta) & \ddot{\eta}'_\theta(\underline{x} - \underline{\lambda}_\theta)\end{bmatrix}$$

$$= E\begin{bmatrix}\dot{\eta}'_\theta(\underline{x} - \underline{\lambda}_\theta)(\underline{x} - \underline{\lambda}_\theta)'\dot{\eta}_\theta & \dot{\eta}'_\theta(\underline{x} - \underline{\lambda}_\theta)(\underline{x} - \underline{\lambda}_\theta)'\ddot{\eta}_\theta \\ \ddot{\eta}'_\theta(\underline{x} - \underline{\lambda}_\theta)(\underline{x} - \underline{\lambda}_\theta)'\dot{\eta}_\theta & \ddot{\eta}'_\theta(\underline{x} - \underline{\lambda}_\theta)(\underline{x} - \underline{\lambda}_\theta)'\ddot{\eta}_\theta\end{bmatrix}$$

$$= \begin{bmatrix} \dot{\eta}_\theta' \Sigma_\theta \dot{\eta}_\theta & \dot{\eta}_\theta' \Sigma_\theta \ddot{\eta}_\theta \\ \ddot{\eta}_\theta' \Sigma_\theta \dot{\eta}_\theta & \ddot{\eta}_\theta' \Sigma_\theta \ddot{\eta}_\theta \end{bmatrix}$$

$$= \begin{bmatrix} \nu_{20}(\theta)=E(\dot{\ell}_\theta^2)=i_\theta & \nu_{11}(\theta)=E_\theta(\dot{\ell}_\theta \ddot{\ell}_\theta)=\text{cov}_\theta(\dot{\ell}_\theta, \ddot{\ell}_\theta) \\ \nu_{11}(\theta)=E_\theta(\ddot{\ell}_\theta \dot{\ell}_\theta)=\text{cov}_\theta(\dot{\ell}_\theta, \ddot{\ell}_\theta) & \nu_{02}(\theta)=E_\theta(\ddot{\ell}_\theta)^2 - i_\theta^2=\text{var}_\theta(\ddot{\ell}_\theta) \end{bmatrix}$$

$$= M_\theta \qquad 2.2.3.15$$

which equals $M_\theta$ as seen earlier and thus we arrived at the statistical definition of curvature which is directly comparable to the geometrical definition of curvature.

## 2.2.4 Some properties of the curvature $\gamma_\theta$

Some of the properties of $\gamma_\theta$ can be summarized:

(1) Statistical curvature is an intrinsic property of the family F and has no dependence on the parametrization. (See p 1195 Efron)

(2) The statistical curvature is invariant under any mapping to a sufficient statistic, including all one-to-one mappings of the sample space. This holds only for the inner product $\Sigma_\theta$ as defined in the definition of statistical curvature.

In order to arrive at a general definition of curvature let

$$F = \{f_\theta(\underline{x}),\ \theta \in \Theta\}$$

be any family of one-parameter densities. The definition of statistical curvature applies as before:-

$$\mathbf{M}_\theta = \begin{bmatrix} \nu_{20}(\theta) & \nu_{11}(\theta) \\ \nu_{11}(\theta) & \nu_{02}(\theta) \end{bmatrix}$$

$$= \begin{bmatrix} E(\dot{\ell}_\theta^2) = i_\theta & E_\theta(\ddot{\ell}_\theta \dot{\ell}_\theta) \\ E_\theta(\dot{\ell}_\theta \ddot{\ell}_\theta) & E_\theta(\ddot{\ell}_\theta)^2 - i_\theta^2 \end{bmatrix} \qquad 2.2.4.1$$

where $\ell_\theta(\underline{x}) \equiv \log f_\theta(\underline{x})$, $\dot{\ell}_\theta(\underline{x}) \equiv \frac{\partial}{\partial\theta} \log f_\theta(\underline{x})$, $\ddot{\ell}_\theta(\underline{x}) \equiv \frac{\partial^2}{\partial\theta^2} \ell_\theta(\underline{x})$

i.e.

$$\gamma_\theta = (|\mathbf{M}_\theta| / i_\theta^3)^{1/2}$$

$$= \left(\frac{i_\theta E_\theta(\ddot{\ell}_\theta^2) - i_\theta^3}{i_\theta^3} - \frac{\nu_{11}(\theta)^2}{i_\theta^3}\right)^{1/2}$$

$$= \left(\frac{E_\theta(\ddot{\ell}_\theta^2) - i_\theta^2}{i_\theta^2} - \frac{\nu_{11}(\theta)^2}{i_\theta^3}\right)^{1/2}$$

$$= \left(\frac{\nu_{02}(\theta)}{i_\theta^2} - \frac{\nu_{11}(\theta)}{i_\theta^3}\right)^{1/2} \qquad 0 < i_\theta < \infty,\ \nu_{02}(\theta) < \infty \qquad 2.2.4.2$$

so that $\gamma_\theta$ measures <u>how quickly Fisher's score statistic is changing as $\theta$ changes</u> and what is more; properties (1) and (2) hold for $\gamma_\theta$ as defined above.

If the arc length from $\eta_{\theta_0}$ to $\eta_\theta$ is written as $s_\theta$ it can be shown that

$$\frac{ds_\theta}{d\theta} = (E_\theta \ell_\theta^2)^{1/2} = i_\theta^{1/2} \qquad 2.2.4.3$$

It can further be shown that if $\eta_\theta$ is large, then the locally best estimator of $\theta$ is changing quickly as $\theta$ changes and F is highly curved in a statistical sense.

These remarks give an indication that the second derivative, curvature as such and a combined study of these quantities may give indications of influential observations as far as curved exponential families are concerned.

A further important fact at this stage is that any smooth one parameter family F can be imbedded in a suitably large exponential family. Suppose that at some point $\theta_0$ in $\Theta$, $\ell_\theta$ is k times differentiable. Consider the k-parameter exponential family

$$g_\eta(\underline{x}) \equiv \exp[\ell_{\theta_0}(\underline{x}) + \eta_1 \dot{\ell}_{\theta_0}(\underline{x}) + \eta_2 \ddot{\ell}_{\theta_0}(\underline{x}) + \ldots + \eta_k \ell_{\theta_0}^{(k)}(\underline{x}) - \varphi(\eta)], \qquad 2.2.4.4$$

$\varphi(\eta)$ being chosen such that the integral over $g_\eta(\underline{x})$ equals 1, where

$$\eta_\theta = [(\theta - \theta_0) \quad \frac{(\theta - \theta_0)^2}{2} \quad \ldots \quad \frac{(\theta - \theta_0)^k}{k!}]$$

so that a one-parameter family of densities $\tilde{f}_\theta \equiv g_{\eta_\theta}$ approximates $f_\theta$ as $\theta \to \theta_0$.

If now the Taylor expansion for $\ell_\theta$ converges at $\theta_0$ the approximation is better

still as $k \to \infty$, i.e. for $k \geq 2$:

$$\tilde{M}_{\theta_0} = M_{\theta_0}, \quad \tilde{i}_{\theta_0} = i_{\theta_0} \text{ and } \tilde{\gamma}_{\theta_0} = \gamma_{\theta_0} . \qquad 2.2.4.5$$

## 2.5 EXTRAPOLATION FROM THE EXPONENTIAL FAMILY TO NON-LINEAR FAMILIES IN GENERAL

Could one expect that results shown in the context of curved exponential families should hold for sufficiently smooth non-exponential families and even non-linear families in general ?

Efron limited himself to the estimators which are smooth functions of the sufficient statistic $\bar{x}$ and which furthermore are consistent and efficient. The estimator of the parameter $\theta$ in a curved exponential family based on an independent identically distributed sample $x_i$, $i=1,\ldots,n$ is determined using the usual least squares or $L_2$-approach. If $\hat{\theta}$ is such an estimator independent of n, $\phi(\theta) \equiv E(\frac{\dot{\ell}_{\theta_0}}{\dot{i}_{\theta_0}} + \theta_0)$ is the best locally unbiased estimator of $\theta$ near $\theta_0$ and $b_\theta \equiv E_\theta(\hat{\theta}) - \theta$ is the bias of $\hat{\theta}$, then

$$\mathrm{Var}_{\theta_0}(\hat{\theta}) = \frac{1}{n\dot{i}_{\theta_0}} + \frac{1}{n^2 i_{\theta_0}} \left\{ \gamma^2_{\theta_0} + 4\frac{\Gamma^2_{\theta_0}}{\dot{i}_{\theta_0}} + \Delta^{\hat{\theta}}_{\theta_0} \right\} + 2\frac{\dot{b}_{\theta_0}}{n\dot{i}_{\theta_0}} + O(\frac{1}{n^2}) \qquad 2.2.5.1$$

where $\Delta^{\hat{\theta}}_{\theta_0} \geq 0$ with equality reserved for the m.l.e. $\hat{\theta}$.

$\Gamma_{\theta_0}$ is the curvature at $\theta = \theta_0$ of the two-dimensional curve $(\theta, \phi(\theta))$ and $\frac{1}{n\dot{i}_{\theta_0}}$

is the Cramér-Rao lower bound for the variance of the unbiased estimator. The quantity $\gamma^2_{\theta_0}$ is the statistical curvature as defined earlier which is invariant under transformations of $\theta$. The quantity $4\frac{\Gamma^2_{\theta_0}}{i_{\theta_2}}$, which Efron calls the "naming curvature" is also known as the "parametric curvature", because its size is directly dependent on how F is parameterized. Under a linear reparameterization however, this value is invariant, i.e. $\theta \rightarrow a + \beta\theta$. The Fisher information is essentially invariant under reparameterizations of F given the acceptance of the usual transformation rule for a differentiable monotomic function $\mu=\mu(\theta)$, i.e. $\underline{i}^T_\mu = \underline{i}^T_\theta(\frac{\partial\theta}{\partial\mu})^2$ for every statistic $T(\underline{x})$. Observe then that $4\frac{\Gamma^2_{\theta_0}}{i_{\theta_0}}$ is not invariant under reparameterization because the solution of the least square problem itself is not invariant under reparameterizations.

It can further be stated that the first term in brackets, i.e. $\gamma^2_{\theta_0}$ is the leading term defining the nonlinearity of a family F, and therefore plays a central role in the calculation of $\lim_{n\to\infty}(\underline{i}_\theta - \underline{i}^{\hat{\theta}}_\theta) = i_\theta\gamma^2_\theta$. In the following chapter we will look at some measures of influence and the curvature will show up again although often in disguised form, i.e. as a factor or term in an expression.

# CHAPTER 3

# RELATIVE CURVATURE MEASURES OF NON-LINEARITY AND MEASURING INFLUENCE IN NON-LINEAR REGRESSION

## 3.1 INTRODUCTION

We observe some measures of influence in linear and non-linear-regression as indicated inter alia by Cook and Weisberg (1982), Belsley et al. (1980) and others as referenced and interpreted by Ross (1984). For the measures of influence in non-linear regression some of these measures may have its usefulness, but other measures are also indicated by Cook and Weisberg (1982) and Atkinson (1985).

Although we had a brief look at curvature and non-linearity in chapter 2 it is of interest to find the relative measures and then also from a differential geometric view point as shown by Ross (1984). An appendix (Appendix A) is given for background reading to differential geometry. The decomposition of the measure of curvature by Bates and Watts (1980) is then used to show some further useful characteristics of relative measures of curvature and its components. Some references are also made to Ratkowsky (1983), Pregibon (1980, 1981) and others.

Observe the model as defined in §2.1.

If $y$:n×1 is the vector of responses and $\eta(\underline{\theta})$:n×1 the vector

$[f(\underset{\sim}{x}_t,\underset{\sim}{\theta})]_{n\times 1}$ , i.e.

$$\underset{\sim}{\eta}(\underset{\sim}{\theta}) = [f(\underset{\sim}{x}_t,\underset{\sim}{\theta})]_{n\times 1} \qquad 3.1.1$$

or

$$\underset{\sim}{\eta}_{\underset{\sim}{\theta}} = \underset{\sim}{f}_{\underset{\sim}{\theta}}$$

then the error vector is

$$\underset{\sim}{e}(\underset{\sim}{\theta}) = \underset{\sim}{y} - \underset{\sim}{\eta}(\underset{\sim}{\theta}) \ , \qquad 3.1.2$$

i.e.

$$S(\underset{\sim}{\theta}) = \underset{\sim}{e}(\underset{\sim}{\theta})'\underset{\sim}{e}(\underset{\sim}{\theta}) = (\underset{\sim}{y} - \underset{\sim}{\eta}(\underset{\sim}{\theta}))'(\underset{\sim}{y} - \underset{\sim}{\eta}(\underset{\sim}{\theta})) \qquad 3.1.3$$

or

$$S_p(\underset{\sim}{\theta}) = \sum_{t=1}^{n} (e_t(\underset{\sim}{\theta}))^p \ , \quad \text{where } e_t(\underset{\sim}{\theta}) \text{ is the t'th element of } \underset{\sim}{e}(\underset{\sim}{\theta}).$$

When $\underset{\sim}{\theta} = \hat{\underset{\sim}{\theta}} \Rightarrow \underset{\sim}{\varepsilon} = \underset{\sim}{e}$ and $\underset{\sim}{\eta}_{\underset{\sim}{\theta}} = \hat{\underset{\sim}{f}}_{\underset{\sim}{\theta}} = \hat{\underset{\sim}{y}}$ .

The expectation surface is defined as

$$\mathbb{M} = \{\underset{\sim}{\eta}(\underset{\sim}{\theta}) : \underset{\sim}{\theta} \in \Omega\}, \qquad 3.1.4$$

where $\mathbb{M} \subset R^n$ and $\Omega \subset R^p$ .

Note that $S(\underset{\sim}{\theta})$ is the squared distance between the observed $\underset{\sim}{y}$ and the point $\underset{\sim}{\eta}(\underset{\sim}{\theta})$ on $\mathbb{M}$, i.e. $\hat{\underset{\sim}{\theta}}$ is the vector of parameters which maps $[f(\underset{\sim}{x}_t,\underset{\sim}{\theta})]_{n\times 1}$ to the point closest to $\underset{\sim}{y}$ on $\mathbb{M}$. Finding $\hat{\underset{\sim}{\theta}}$ involves firstly determining the nearest point in

the expectation surface $\mathbf{M}$ to $\underline{y}$ and secondly finding the corresponding $\hat{\underline{\theta}}$.

As before assume that $\underline{\varepsilon} \sim n(\underline{0}, \sigma^2 I)$. Because of this last assumption it follows that the least squares estimates of $\underline{\theta}$, i.e. $\hat{\underline{\theta}}$ are also the maximum likelihood estimators - although this is not necessarily true for the $L_p$-estimators when $p \neq 2$. (The p here not to be confused with p, the number of parameters in the non-linear function. The distinction should be obvious from the context).

If $\eta_{\underline{\theta}}$ is non-linear in the parameters and $\mathbf{M}$ is not a linear subspace of $\mathbb{R}^n$, then none of the properties usually assumed for the linear alternative hold, i.e.

$$\hat{\underline{\theta}} \sim n_p(\underline{\theta}, (X'X)^{-1}\sigma^2)$$

$$|\underline{y} - \hat{\underline{y}}|^2/\sigma^2 \sim \chi^2_p$$

$$|\hat{\underline{e}}|^2/\sigma^2 \sim \chi^2_{n-p} \qquad 3.1.5$$

$\hat{\underline{\theta}}$ and $s^2$ are independent where $s^2 = \dfrac{S(\hat{\underline{\theta}})}{n - p} = \dfrac{\underline{e}'\underline{e}}{n - p}$

Statistics designed to assess the effect of an individual case or group of cases on some aspect(s) of the subsequent analysis are termed influence measures. A widely used technique for generating such statistics is the method of case deletion where an individual case or group of cases is removed and the regression is repeated on the remaining data. Statistics are then constructed which reflect the important differences in the resulting analysis from that obtained by regression on the full data set.

## 3.2 MEASURES OF INFLUENCE IN THE LINEAR MODEL

### 3.2.1 Measures based on the hat matrix and the elliptical norm

Let J be a subset of the integers $\{1,\ldots,n\}$. Indicate the least squares estimate of $\underline{\theta}$ obtained after deleting all cases whose indices are in J by $\hat{\underline{\theta}}_{(J)}$. As $\hat{\underline{\theta}}_{(J)} - \hat{\underline{\theta}}$ has no natural ordering as a vector in $R^p$, a mapping must be constructed in $R^1$ which will allow the investigator to assess the magnitude of $\hat{\underline{\theta}}_{(J)} - \hat{\underline{\theta}}$ in a meaningful way. Some influence functions derived with this in mind are based on the elliptical norms in the following way:-

Let (A,c) define an influence measure as follows:

$$D_J(A,c) = (\hat{\underline{\theta}}_{(J)} - \hat{\underline{\theta}})'A(\hat{\underline{\theta}}_{(J)} - \hat{\underline{\theta}})/c \qquad 3.2.1.1$$

where $|A| \geq 0$, A:p×p and symmetric and $c > 0$ constant, A and c both being dependent on the aspect(s) under analysis. For example Cook's statistic is defined in general as

$$D_J(X'X,ps^2) = (\hat{\underline{\theta}}_{(J)} - \hat{\underline{\theta}})'(X'X)(\hat{\underline{\theta}}_{(J)} - \hat{\underline{\theta}})/ps^2 \qquad 3.2.1.2$$

for the linear model $\underline{Y} = X\beta + \underline{\varepsilon}$.

In the case of linear regression with normally distributed errors a $(1 - \alpha)\times 100\%$ confidence region is given by the elliptical region

$$\{\underline{\theta}:(\hat{\underline{\theta}}_{(J)} - \hat{\underline{\theta}})'\ (X'X)(\hat{\underline{\theta}}_{(J)} - \hat{\underline{\theta}}) \leq ps^2 f_{p,n-p,\alpha}\} \qquad 3.2.1.3$$

where $f_{p,n-p,\alpha}$ is the upper $\alpha$-percentage point of the $F_{p,n-p}$ distribution. Now, Cook's distance can be compared with the percentage points of the $F_{p,n-p}$ distribution; because although $D_J(X'X,ps^2)$ is not distributed $F_{p,n-p}$, we find a direct analogy which gives a convenient approximate scale with which one can assess the effect of removing those cases designated by J. One can take as an example $J = \{i\}$, then $D_i = F(p,n - p;0,50)$ and the simultaneous deletion of $\underline{x}_i$ moves the estimate of $\underline{\theta}$ to the edge of a 50% confidence ellipsoid relative to $\underline{\theta}$. Note that $h_i/(1 - h_i) = \underline{x}_i'(X_{(i)}'X_{(i)})^{-1}\underline{x}_i$ is called the potential $p_i(X'X)$, where $h_i$ is the leverage of observation i, i.e. the ith diagonal element of the hat matrix $H = X(X'X)^{-1}X'$.

Cook's distance can be written in several forms, viz

$$D_J = |\underline{\eta}(\underline{\hat{\theta}}_{(J)}) - \underline{\hat{\eta}}|/ps^2 \qquad 3.2.1.4$$

which measures the stability of the fitted values with respect to the indexed cases, or

$$D_J = [S(\underline{\hat{\theta}}_{(J)}) - S(\underline{\hat{\theta}})]/pS(\underline{\hat{\theta}}) \qquad 3.2.1.5$$

where $S(\underline{\hat{\theta}})$ is a measure of the precision of fit to the data by the model $\underline{\eta}(\underline{\theta})$, so that $D_J$ is proportional to the relative change in precision resulting from deletion of those cases represented by J and thus reflects the stability of the fit to the full data set with respect to the J deleted cases, and

$$D_i = r_i^2 h_i/[p(1 - h_i)] \qquad 3.2.1.6$$

for the single case deletion with J = {i}. Note that

$$r_i = e_i/[s\sqrt{1 - h_i}] \ , \quad i = 1,\ldots,n \qquad 3.2.1.7$$

is known as the standardized residual according to Cook and Weisberg (1982).

If $\underline{e} \sim n(\underline{0},\sigma^2(I - H))$ then $var(r_i) = 1$, $\underline{e} = \underline{y} - \hat{\underline{y}} = (I - H)\underline{y}$ and the $r_i$'s are not independent.

Kuh and Welsch (1977) used the form

$$D_J(X'X,ps^2_{(J)}) = (\hat{\underline{\theta}}_{(J)} - \hat{\underline{\theta}})'X'X(\hat{\underline{\theta}}_{(J)} - \hat{\underline{\theta}})/ps^2_{(J)} \qquad 3.2.1.8$$

where $s^2_{(i)} = s^2 \dfrac{n - p - r_i^2}{n - p - 1}$ when J = {i} so that for J = {i}

$$D_i = t_i^2 h_i/[p(1 - h_i)] \qquad 3.2.1.9$$

with $t_i$ the externally studentized residual

$$t_i = e_i/[s_{(i)}\sqrt{1 - h_i}] \ , \quad i = 1,\ldots,n \qquad 3.2.1.10$$

which follows Student's t-distribution with n - p - 1 degrees of freedom. This measure of Kuh and Welsch for a single case deletion is called $DFFITS_i$ by Belsley et al. (1980).

The difference between Cook's measure and the measure of Kuh and Welsch is that

in the former case comparisons of the measures between cases are possible because of the constant (similar) norm (scaling factor) used. In the latter case the norm or scaling factor changes for every new combination of case(s) deleted, although the fact that $t_i^2 \sim F_{1,n-p-1}$ has some advantages, e.g. in the construction of formal test measures.

In the hat matrix the element $h_i = h_{ii}$ is called the leverage component with respect to the i'th case. Hoaglin and Welsch (1978) discussed the uses of $H = [h_{ij}]$ where H is symmetric and idempotent with some characteristics such as

$$\sum_{i \neq j} h_{ij}^2 = h_i(1 - h_i) \qquad i = 1,\ldots,n$$

$$\Rightarrow 0 \leq h_i \leq 1$$

and

$$\begin{aligned} tr(H) &= tr(X(X'X)^{-1}X') \\ &= tr\{X'X)^{-1}X'X\} \\ &= tr(I) \\ &= p \end{aligned}$$

$$\text{i.e. } \sum_{i=1}^{p} h_i = 1 \qquad 3.2.1.11$$

If $\hat{y}_i$ is the i'th fitted value, then

$$\hat{y}_i = h_i y_i + \sum_{j \neq i} h_{ij} y_j \qquad i = 1,\ldots,n \qquad 3.2.1.12$$

so that if $h_i = 1 \Rightarrow \hat{y}_i = y_i$ fits exactly.

As $h_i$ diminishes, the influence measure will decrease, i.e. $h_i \rightarrow 0$ leads to

$D_i \rightarrow 0$; This means small leverage, small influence, unless however $r_i$ or $t_i$ is large which indicates a value which appears to be discrepant in the same sense as outliers. Cook and Weisberg (1982) showed that cases remote in the design space have large leverage so that $var(e_i) = (1 - h_i)\sigma^2$ is small making outlier detection difficult. This is the reason why $r_i$ and $t_i$ have the denominators $s(1 - h_i)$ and $s_{(i)}(1 - h_i)$ respectively.

### 3.2.2 The volume ratio measures, the Kullback-Leibler measure and the measure of Andrews and Pregibon

Cook and Weisberg (1982) use the ratio of the volumes of the ellipsoids resulting from the perturbed data matrix to that of the original data matrix as another measure of influence, i.e.

$$VR_i = -\frac{1}{2}\ell og(1 - h_i) - \frac{p}{2}\ell og\frac{n - p - 1}{n - p - r_i^2}\frac{F_p}{F_p^1}$$

where $F_p = F_{p,n-p;1-\alpha}$ and $F_p^1 = F_{p,n-p-1;1-\alpha}$. This is the form of the measure if a y-intercept is present. If there is no intercept term then

$$VR_{i(\text{no intercept})} = VR_i + \frac{1}{2}\ell og(1 - \frac{1}{n}). \qquad 3.2.2.1$$

A negative measure would indicate that the observation deleted resulted in a decrease of volume, i.e. in an increase in precision. This will happen if $r_i^2$ is large and the corresponding leverage, $h_i$ small. The opposite is true for a positive measure which implies a volume increase and hence a decrease in precision. This is true for large leverage, i.e. $h_i$.

A further measure is of a Bayesian type, i.e. the Kullback-Leibler measure which can be seen as a prior influence function(PIF). Cook and Weisberg give the Kullback-Leibler measure for $\sigma^2$ known, an approximation for the case when $\sigma^2$ is unknown, an analytical form for m = 1 as well as a predictive form of the measure. $H_J$ will refer to the determination of H with the deletion of the subset J of the observation vectors. Note that m = #(J) is the number of probable influential vectors.

For $\sigma^2$ known the Kullback-Leibler measure of influence is - see 3.2.1.8

$$d_J = D_J(X'X,4\sigma^2) - \tfrac{1}{2}\ell\text{og}|I + \tfrac{1}{2}H_J(I - H_J)^{-1}| + \tfrac{1}{4}\text{tr}[H_J(I - H_J)^{-1}] \qquad 3.2.2.2$$

(i.e. the change in centre - the change in volume + ratio of sum of variances)

An approximation for $d_J$ when $\sigma^2$ unknown is as follows

$$\tilde{d}_J = \frac{n - p - 2}{n - p}D_J(X'X,4\hat{\sigma}^2) - \tfrac{1}{2}\ell\text{og}|I + \tfrac{1}{2}H_J(I - H_J)^{-1}| +$$

$$\tfrac{1}{4}k_J\text{tr}[H_J(I - H_J)^{-1}] + \frac{n}{2}[k_J - \ell\text{og } k_J - 1] \qquad 3.2.2.3$$

where $k_J = \frac{n - p - 2}{n - m - p - 2}\left(1 - \frac{r_J^2}{n - p}\right)$

If m = 1, say J = {i},

$$\tilde{d}_i = \frac{p(n - p - 2)}{4(n - p)}D_i + \frac{k_i}{4}\left(\frac{h_{ii}}{1 - h_{ii}}\right) - \frac{1}{2}\ell\text{og}\left(1 + \frac{\frac{1}{2}h_{ii}}{1 - h_{ii}}\right) + \frac{n}{2}(k_i - \ell\text{og } k_i - 1)$$

3.2.2.4

When a predictive (future) matrix X, say $X_f$ is applicable $\tilde{d}_i(\underline{x}_f)$ must be calculated for every $\underline{x}_f \in X_f$ and the maximum of the $\tilde{d}_i(\underline{x}_f)$'s taken over all the $\underline{x}_f$'s is being employed as the influence of $X_f$, i.e.

$$\tilde{d}_i^* = \max_{\underline{x}_f}\{\tilde{d}_i(\underline{x}_f)\} \qquad 3.2.2.5$$

It can be shown that

$$2\tilde{d}_i^* = \frac{n-p-2}{n-p}r_i^2\,\frac{h_{ii}-\frac{1}{n+1}}{1-h_{ii}} + k_i\,\frac{h_{ii}-\frac{1}{n+1}}{1-h_{ii}} -$$

$$\ell og\left\{1+\frac{h_{ii}-\frac{1}{n+1}}{1-h_{ii}}\right\} + k_i - \ell ogk_i - 1 \qquad 3.2.2.6$$

If we define

$$R_J = (1-\frac{r_J^2}{n-p})|I - H_J|$$

then a last measure as introduced by Andrews and Pregibon (1978) is inter alia based on the change in the residual sum of squares as well as $|X'X|$. The measure is unitless and the ratio of ellipsoidal volumes is again present:

$$AP_J = -\frac{1}{2}\ell ogR_J = -\frac{1}{2}\ell og|I - H_J| + \frac{1}{2}\ell og\frac{n-p}{n-p-r_J^2} \quad (\text{See } VR_i) \qquad 3.2.2.7$$

This measure is large for influential cases. Unfortunately it poses some formidable computational problems (Cook and Weisberg, 1982).

## 3.3 FURTHER ASPECTS OF CURVATURE IN NON-LINEAR REGRESSION MODELS

### 3.3.1 The connection coefficients and coefficients of the second fundamental form

In §2.2 the statistical/geometrical curvature approach of Efron was discussed. In the multi-parameter situation however, it is easier to deal with the differential geometric approach of Amari (1982) and the lifted line approach of Bates and Watts (1980). For more information on differential geometry refer to appendix A.

In order to find measures of curvature in terms of differential geometry we need some definitions:-

Let $\underline{\eta}(\underline{\theta})$ be the vector function which defines the expectation surface $\mathbf{M}$. The partial derivatives with respect to the components of $\underline{\theta}$, i.e. $\theta_i$, $i = 1,\ldots,p$ are shown as follows

$$\frac{\partial \underline{\eta}(\underline{\theta})}{\partial \theta_i} = \underline{\eta}_i \ , \quad \frac{\partial^2 \underline{\eta}(\underline{\theta})}{\partial \theta_i \partial \theta_j} = \underline{\eta}_{ij} \qquad 1 \leq i,j \leq p \qquad 3.3.1.1$$

Given a fixed value of $\underline{\theta}$, say $\underline{\theta}_0$, then the tangent space to $\mathbf{M}$ at the point $\underline{\eta}(\underline{\theta}_0)$ is the subspace of $\mathbf{R}^n$ spanned by the vectors $\underline{\eta}_i$ with $i = 1,\ldots,p$ and evaluated at $\underline{\theta}_0$. The tangent plane to $\mathbf{M}$ at the point $\underline{\eta}(\underline{\theta}_0)$ is the tangent space at $\underline{\eta}(\underline{\theta}_0)$ translated to $\underline{\eta}(\underline{\theta}_0)$, see Ross (1984).

The Einstein Summation convention will be used in order to simplify expressions

containing summations in expressions. An index being repeated as sub- and superscript indicates a summation over that index. Take for example the vectors $\underline{\eta}_i$, i = 1,...,p evaluated at $\underline{\theta}_0$ which form a basis for the tangent space of $\mathbf{M}$ at $\underline{\eta}(\underline{\theta}_0)$. If $\underline{\nu}$ is an element of that space it may be written as

$$\underline{\nu} = \sum_{i=1}^{p} \nu_i \underline{\eta}_i \text{ or as } \underline{\nu} = \nu^i \underline{\eta}_i \ . \qquad 3.3.1.2$$

If one has a non-linear function $\underline{\eta}(\underline{\theta})$ which defines $\mathbf{M}$, then the non-linearity of $\underline{\eta}(\underline{\theta})$ may be such that the non-linearity corresponds to <u>intrinsically linear</u> modelling where the non-linearity is a case of parameterization, i.e. if $\underline{\theta} = \underline{\theta}(\underline{\beta})$ then $\underline{\eta}(\underline{\theta}(\underline{\beta})) = \mathbf{X}\underline{\beta}$, where $\mathbf{X}$ is a matrix dependent on the experimental settings only. In such a case $\mathbf{M}$ is a planar subspace of $\mathbb{R}^n$ spanned by the columns of $\mathbf{X}$.

On the other hand the non-linearity $\underline{\eta}(\underline{\theta})$ may be the result of the parameterization as well as the intrinsic non-linearity.

Beale (1960) and Bates and Watts (1980) discussed some measures of non-linearity in terms of differential geometry.

In order to decompose the second partial derivatives $\underline{\eta}_{ij}$ of the expectation surface $\mathbf{M}$ into their tangent and normal components, $\underline{\eta}^T_{ij}$ and $\underline{\eta}^N_{ij}$, see §3.3.2 together with Amari (1982). Ross (1984) called the subspace spanned by the normal components $\underline{\eta}^N_{ij}$ evaluated at $\underline{\theta}_0$ the acceleration space of $\mathbf{M}$ at $\underline{\eta}(\underline{\theta}_0)$ and denoted it by $(E_1)_{\underline{\eta}(\underline{\theta}_0)}$. If dimension$(E_1)$ = m, construct the orthonormal basis for $(E_1)_{\underline{\eta}(\underline{\theta}_0)}$, the acceleration space, i.e. $\underline{\xi}_i$, i = 1,...,m, so that

$$\eta_{ij} = \Gamma^{k}_{ij}\eta_{k} + b^{a}_{ij}\xi_{a} \qquad 1 \leq i,j \leq p \qquad 3.3.1.3$$

with $\Gamma^{k}_{ij}$ , k = 1,...,p being obtained from the solution of the regression of $\eta_{ij}$ on the first partial derivatives - $\Gamma^{k}_{ij}$ being the Christoffel symbols of the second kind, or the connection coefficients. Similarly $b^{a}_{ij}$, a = 1,...,m can be obtained from the regression of $\eta_{ij}$ on the orthonormal basis $\xi_{a}$, a = 1,...,m of the acceleration space. They are called the coefficients of the second fundamental form of M.

In order to evaluate non-linearity it is of some importance to observe the implications of a parameter transformation.

If $\underline{\theta} = \underline{\theta}(\underline{\beta})$ is a parameter transformation, let $\bar{\eta}(\underline{\beta}) = \eta(\underline{\theta}(\underline{\beta}))$, so that

$$\bar{\eta}_{uv} = \Gamma^{t}_{uv}\bar{\eta}_{t} + b^{a}_{uv}\,\xi_{a} \qquad 3.3.1.4$$

where the right hand side terms are the tangential and normal components respectively.

Ross showed that both $\Gamma^{k}_{ij}$ and $b^{a}_{ij}$ provide information about the non-linearity of M. If $b^{a}_{ij}$ are all zero for some parameterization, then they must be zero for every parameterization. The resulting argument then is that M is a plane since $b^{a}_{ij}$ are constructed from the normal components of the second order partial derivatives. (see Appendix A) It follows further that if the transformation is linear then the connection coefficients $\Gamma^{k}_{ij}$ will also be zero. If however the transformation is non-linear, the connection coefficients will not be zero in general. This shows that if the connection coefficients as well as the

coefficients of the second fundamental form are zero then $\eta_{ij} = 0$ and $\underline{\eta}$ is linear in the parameters - refer also Appendix A.5. Ross then made use of these facts to derive some measures of non-linearity:-

Define the first derivatives as follows

$$V = [v_{ij}]_{n\times p}$$

$$= \begin{bmatrix} \frac{\partial \eta_1(\theta)}{\partial \theta_1} & \cdots\cdots & \frac{\partial \eta_1(\theta)}{\partial \theta_p} \\ \cdots\cdots & \cdots\cdots & \cdots\cdots \\ \frac{\partial \eta_n(\theta)}{\partial \theta_1} & \cdots\cdots & \frac{\partial \eta_n(\theta)}{\partial \theta_p} \end{bmatrix}$$

$$= [\eta_1, \eta_2, \ldots, \eta_p] \qquad 3.3.1.5$$

then $G = V'V = [g_{ij}]_{p\times p}$ and $G^{-1} = [g^{ij}]$. If $\underline{v} = v^i \eta_i$ is an element of the tangent space of $M$ at $\eta(\underline{\theta})$, then $\underline{v}'\underline{v} = v^i \eta_i \, v^{ji} \eta_{ji} = v^i v^j g_{ij}$, so that $\eta_i' \eta_j = g_{ij}$.

Now we can define:

$$\Gamma_{ijk} = \Gamma_{ij}^{\ell} g_{\ell k}$$

$$b_{ija} = b_{ij}^{c} \xi_c' \xi_a = b_{ij}^{a} \qquad 3.3.1.6$$

[e.g. $\underline{n}_{ij}' \eta_{k\ell} = (\Gamma_{ij}^{k} \eta_k + b_{ij}^{a} \xi_a)' (\Gamma_{k\ell}^{r} \eta_r + b_{k\ell}^{s} \xi_s)$

$$= \Gamma^k_{ij}\Gamma^r_{k\ell}\eta_k\eta_r + b^a_{ij}b_{k\ell a}$$

$$= \Gamma^k_{ij}\Gamma_{k\ell r} + b^a_{ij}b_{k\ell a} \,]$$

Given $B^a = (b^a_{ij})$ and $K^a = G^{-1}B^a$ some measures of curvature are defined:

- The squared length of the second fundamental form, i.e.

$$\nu = \sum_a (\mathrm{tr}(K^a)^2) \geq 0 \text{ in general}$$

$$= 0 \text{ if } M \text{ is a plane} \qquad 3.3.1.7$$

- The mean curvature vector:

$$\rho = \frac{1}{p}(\mathrm{tr}(K^a))\xi_a \text{ with} \qquad 3.3.1.8$$

$|\rho|^2 = \frac{1}{p^2}\sum_a (\mathrm{tr}(K^a))^2 = 0$ if $M$ is a plane, although the mere fact that $|\rho|^2 = 0$ does not necessarily imply that $M$ is planar.

- The scalar curvature:

$$\chi = p^2\,|\rho|^2 - \nu$$

$$= \sum_a (\mathrm{tr}(K^a))^2 - \sum_a (\mathrm{tr}(K^a)^2) \qquad 3.3.1.9$$

which is a function of the coefficients of the second fundamental form. If now $M$ is a nonplanar expectation surface, then it is possible to find a parameterization such that the connection coefficients are all zero, but only if $\chi = 0$ on $M$, i.e. only for those models generating an expectation surface having zero scalar curvature, which means that it is not true in general and thus very restrictive.

### 3.3.2 The normal and tangential components of curvature

The lifted line measures of Bates and Watts (1980) can briefly be described as follows:

Given

$$\underline{\theta}(t) = \underline{\theta}_0 + \underline{\delta}t \qquad 3.3.2.1$$

where $\underline{\delta}_{p\times 1}$ is an arbitrary non-zero vector and $\underline{\theta}_0$ a constant parameter vector, the curve $\underline{c}_{\underline{\delta}}(t) = \eta(\underline{\theta}(t))$ is called a lifted line. As the straight line $\underline{\theta}(t)$ in $\Omega$ passes through $\underline{\theta}_0$ in the direction $\underline{\delta}$, $\underline{c}_{\underline{\delta}}(t)$ will form a curve in $\mathbb{R}^n$ on $M$ and the non-linearity of $\eta(\underline{\theta})$ will be shown by the properties of the curves $\underline{c}_{\underline{\delta}}(t)$ for all directions $\underline{\delta}$. Each of these curves is known as the "lifted line" of Bates and Watts for different values of the parameter t or parameter vector $\underline{\theta}_0$.

Let $\dot{\underline{c}}_{\underline{\delta}} = \dot{\underline{c}}_{\underline{\delta}}(t)|_{t=0}$ be the tangent to the curve $\underline{c}_{\underline{\delta}}$ at $t = 0$ and $\underline{\theta} = \underline{\theta}_0$ and $\ddot{\underline{c}}_{\underline{\delta}} = \ddot{\underline{c}}_{\underline{\delta}}(t)|_{t=0}$ where derivatives are taken with respect to t, i.e.

$$\dot{\underline{c}}_{\underline{\delta}} = \frac{d\underline{c}_{\underline{\delta}}}{dt}\Big|_{t=0} = \sum_i \left(\frac{\partial \underline{c}}{\partial \theta_i}\frac{d\theta_i}{dt}\right)_{t=0} \qquad 3.3.2.2$$

$$\ddot{\underline{c}}_{\underline{\delta}} = \frac{d^2\underline{c}_{\underline{\delta}}}{dt^2}\Big|_{t=0} = \sum_j \left(\frac{\partial \dot{\underline{c}}_{\underline{\delta}}}{\partial \theta_j}\frac{\partial \theta_j}{\partial t}\right)_{t=0}$$

$$= \ddot{c}^{T}_{\delta} + \ddot{c}^{N}_{\delta}$$

$$= \ddot{c}^{P}_{\delta} + \ddot{c}^{G}_{\delta} + \ddot{c}^{N}_{\delta} \qquad 3.3.2.3$$

where the first two terms on the right hand side refer to the tangential component and the third term to the normal component. Whereas $\ddot{c}^{P}_{\delta}$ refers to the change in speed of the moving point and thus to the uniform motion - when applicable - of the moving point across the solution locus, the geodesic acceleration $\ddot{c}^{G}_{\delta}$ determines the change in direction of the vector $\dot{c}_{\delta}$ parallel to the tangent plane. The normal acceleration $\ddot{c}^{N}_{\delta}$ however, determines the change in direction of the vector $\dot{c}_{\delta}$ normal to the tangent plane. Then the maximum intrinsic curvature of M at $\eta(\theta_0)$ is defined from the right hand side by

$$k^{N} = \sup_{\delta} \{k^{N}_{\delta}\} = \sup_{\delta}\{|\ddot{c}^{N}_{\delta}|/|\dot{c}_{\delta}|^{2}\} \qquad 3.3.2.4$$

which is identical to the first Frenet curvature of M in differential geometry (Thomas, 1961). Note that the intrinsic curvature of the solution locus is a special case of Efron's statistical curvature as discussed in chapter 2. (Bates and Watts, 1981 and Reid in Bates and Watts, 1980). Further

$$k^{T} = \sup_{\delta} \{k^{T}_{\delta}\} = \sup_{\delta}\{|\ddot{c}^{T}_{\delta}|/|\dot{c}_{\delta}|^{2}\} \qquad 3.3.2.5$$

is defined as the maximum parameters effect curvature. Ratkowsky (1983) described the parameters effect curvature as an indication of the lack of

parallelism and the unequal spacing of the projection of the parameter contours on the tangent plane.

These measures can be examined in terms of the coefficients of acceleration or second fundamental form and tangency or connection by computing explicitly the vectors $\dot{\underline{c}}_{\underline{\delta}}$ and $\ddot{\underline{c}}_{\underline{\delta}}$ as follows:

From $\underline{c}_{\underline{\delta}}(t) = \underline{\eta}(\underline{\theta}(t)) = \underline{\eta}(\underline{\theta}_0 + \underline{\delta}t) \qquad (\propto \underline{\eta}(\underline{\delta}_t))$

$$\dot{\underline{c}}_{\underline{\delta}} = \delta^i \eta_i$$

$$\ddot{\underline{c}}_{\underline{\delta}} = \delta^i \delta^j \eta_{ij}$$

$$= \delta^i \delta^j \Gamma^k_{ij} \eta_k + \delta^i \delta^j b^a_{ij} \xi_a \qquad 3.3.2.6$$

being the sum of the tangential and acceleration components. It now follows that

$$|\ddot{\underline{c}}^T_{\underline{\delta}}|^2 = \Gamma^r_{ij} \Gamma_{k\ell r} \delta^i \delta^j \delta^k \delta^\ell$$

$$|\ddot{\underline{c}}^N_{\underline{\delta}}|^2 = b^a_{ij} b_{k\ell a} \delta^i \delta^j \delta^k \delta^\ell$$

$$|\dot{\underline{c}}_{\underline{\delta}}|^2 = g_{ij} \delta^i \delta^j \qquad 3.3.2.7$$

and as $\bar{\underline{\delta}} = \underline{\delta}/|\underline{c}_{\underline{\delta}}|^2$ yields

$$k^N_{\underline{\delta}} = \sqrt{b^a_{ij} b_{k\ell a} \bar{\delta}^i \bar{\delta}^j \bar{\delta}^k \bar{\delta}^\ell}$$

$$k^T_{\underline{\delta}} = \sqrt{\Gamma^r_{ij} \Gamma_{k\ell r} \bar{\delta}^i \bar{\delta}^j \bar{\delta}^k \bar{\delta}^\ell}$$

where $|\dot{\underline{c}}_{\underline{\delta}}|^2 = 1$,

$$k^N_{\underline{\delta}} = k^N_{\underline{\delta}} \text{ and } k^T_{\underline{\delta}} = k^T_{\underline{\delta}} \,. \qquad 3.3.2.8$$

So again the maxima of $k^T_{\underline{\delta}}$ and $k^N_{\underline{\delta}}$ on $\{\underline{\delta}: g_{ij}\delta^i\delta^j = 1\}$ give the <u>parameters effect curvature</u> and <u>intrinsic curvature</u>. As both functions are continuous, the maxima do exist.

The measures up to now are absolute in nature. In order to be able to assess whether the maximum curvature, i.e. intrinsic or parameters effect, will affect inferences based on linear approximations one must determine the relative curvature measures.

Given further the orthogonal vector to the tangent space of M at a given point and call this vector $\xi$, then following Ross (1984):

$$k^N_{\xi,\underline{\delta}} = |\xi' \ddot{\underline{c}}_{\underline{\delta}}|/(|\xi||\dot{\underline{c}}_{\underline{\delta}}|^2) \qquad 3.3.2.9$$

which is the <u>normal</u> curvature of $\underline{c}_{\underline{\delta}}$ relative to the direction $\xi$. Similarly, the relative intrinsic curvature of M is then

$$k^N_{\xi} = \sup_{\underline{\delta}}\{k^N_{\xi,\underline{\delta}}\} \leq k^N_{\underline{\delta}} \qquad 3.3.2.10$$

with equality for $k^N_{\underline{\delta}} = 0$ or if $\xi \propto \underline{c}^N_{\underline{\delta}}$ .

Let

$$B_{\xi} = [\xi' \eta_{ij}], \qquad 3.3.2.11$$

then

$$|\xi| k^N_{\xi,\underline{\delta}} = \frac{|\xi||\xi' \underline{\ddot{c}}_{\underline{\delta}}|}{|\xi||\underline{\dot{c}}_{\underline{\delta}}|^2}$$

$$= \frac{|\xi||\xi' \delta^i \delta^j \eta_{ij}|}{|\xi||g_{ij}\ \delta^i \delta^j|}$$

$$= |\underline{\delta}'\ B_{\xi}\ \underline{\delta}|/\underline{\delta}' G \underline{\delta} \qquad 3.3.2.12$$

from 3.3.2.6 and 3.3.2.7.

Then it follows that if

$$\lambda_{\xi} = \underset{\substack{\text{absolute}\\ \text{eigenvalue}}}{\max}(G^{-1} B_{\xi}) \qquad 3.3.2.13$$

then the relative <u>intrinsic</u> curvature of M in the direction $\xi$ is

$$k^N_{\xi} = |\lambda_{\xi}|/|\xi| \qquad 3.3.2.14$$

Separating into intrinsic and parameters effect curvature is important as a linear approximation combines a planar assumption together with a uniform coordinate assumption (Bates and Watts, 1980). The first involves the replacement of the curved solution by a tangent plane which is only acceptable if the maximum locus intrinsic curvature is small at $\eta(\hat{\theta})$. The latter assumption implies that the curved parameter lines on this approximating tangent plane is replaced by a grid of straight, parallel, equispaced lines which will only be acceptable if the maximum parameters effect curvature is small at $\eta(\hat{\underline{\theta}})$. It is important to note that the planar assumption must be satisfied before one may carry on further investigations by means of methods which make use of the approximate linearity assumption.

Bates en Watts (1980) suggest that linear approximations should be adequate if the maximum parameter-effects and maximum intrinsic curvatures are both small compared to the guide

$$c = \{F_{\alpha}(p,n - p\}^{-1/2} \qquad 3.3.2.15$$

but Ratkowsky (1983) proposed a cut-off level of c/2.

Cook and Goldberg (1986) however, demonstrate in the Fieller-Creasy problem that the Bates and Watts procedure can fail for subsets of $\underline{\theta}$ where the ratio of the means of two normal populations are present.

Suppose now the planar approximation is not adequate enough, i.e. the non-linear function may not be replaced by a linear approximation, how would we determine an influential observation ? According to Donaldson and Schnabel (1986) there

are more than enough cases where the linearization method give poor results indeed in spite of claims in many instances where this in not the case. They found that the Bates and Watts results are far more reliable in non-linear situations than the linear approximation procedures.

## 3.4 MEASURES OF INFLUENCE IN THE NON-LINEAR MODEL

Cook and Weisberg (1982) suggested that the general diagnostic methods derived for linear models may be applied almost directly to non-linear models. A mayor change is being implemented in terms of the level of the diagnostic tool used. See also Fox et al (1984). In stead of using the data matrix as such, the jacobian of the model may be determined and substituted for the data matrix. Hoaglin and Velsch (1978) stated that the diagonal elements $h_i$ of $\hat{J}(\hat{J}'\hat{J})^{-1}\hat{J}'$ may serve as diagnostics to identify high leverage points. They motivate their approach as follows:

Let

$$y = f(\underline{x},\underline{\theta}) \text{ and } F = \Sigma(y_i - f(\underline{x}_i,\underline{\theta}))^2 \qquad 3.4.1$$

and assume approximate linearity in a neighbourhood about $\hat{\underline{\theta}}$, i.e.

$$f(\underline{x}_i,\underline{\theta}) \simeq f(\underline{x}_i,\hat{\underline{\theta}}) + \hat{j}_i'(\underline{\theta} - \hat{\underline{\theta}}) \qquad 3.4.2$$

or rather

$$\underline{f}(\underline{x},\underline{\theta}) \simeq \underline{f}(\underline{x}_i,\hat{\underline{\theta}}) + \hat{J}'(\underline{\theta} - \hat{\underline{\theta}}) \qquad 3.4.3$$

where $\hat{j}_i'$ is the i-th row of the n×p Jacobian matrix $\hat{J}$ as defined in chapter 2.1. If we indicate the one-step estimators of $\underline{\hat{\theta}}_{(i)}$ by $\underline{\hat{\theta}}^1_{(i)}$ in the minimization of

$$F_{(i)} = \sum_{j \neq i} (y_j - f(\underline{x}_j, \underline{\theta}))^2, \qquad i = 1,\ldots,n \qquad 3.4.4$$

and substitute 3.4.2 in 3.4.4 then the new function obtained is minimized at

$$\underline{\theta}^1_{(i)} = \underline{\hat{\theta}} + (\hat{J}'_{(i)}\hat{J}_{(i)})^{-1}\hat{J}'_{(i)}\underline{e}_{(i)} \qquad 3.4.5$$

with $\underline{e}_{(i)}$ having the usual meaning. This estimator is similar to that obtained using a single step of the Gauss-Newton method as indicated in chapter two. It can now be shown that (Appendix C)

$$\underline{\hat{\theta}}^1_{(i)} = \underline{\hat{\theta}} - \frac{(\hat{J}'\hat{J})^{-1}j_i e_i}{1 - \hat{h}_i} \qquad 3.4.6$$

with $\hat{h}_i$ as defined in 3.2.1, but in terms of the Jacobian i.e.

$$\hat{h}_i = \hat{j}_i'(\hat{J}'\hat{J})^{-1}\hat{j}_i. \qquad 3.4.7$$

The difference between $\underline{\hat{\theta}}^1_{(i)}$ and $\underline{\hat{\theta}}$ can now be applied in Cook's statistic or any other appropriate analysis in order to determine influence - e.g. the measures as defined in §3.2 with J as norm in stead of X and variations thereof as indicated by Atkinson (1985).

The underlying assumption of approximate linearity in the neighbourhood of the

estimator plays a crucial role. When this assumption is true "enough" according to the measures of non-linearity, say relative curvature, as discussed in §3.3 then the diagnostics and residual analysis for linear least squares may be expected to apply at least approximately in non-linear least squares.

Should $f(\underset{\sim}{x},\underset{\sim}{\theta})$ have markedly non-elliptical contours as for some non-linear models the use of elliptical norms for diagnostics and residual analysis may be inappropriate whether based on one-step or fully iterated schemes. Not only will the non-ellipsoidicity result in poor estimates of the effect on parameter estimation using the deletion of the i-th observation through $\hat{\underset{\sim}{\theta}}_{(i)}$, but also the statistics of Cook or any of its variations may provide poor estimates of the effect on the residual sum of squares moving from $\hat{\underset{\sim}{\theta}}$ to $\hat{\underset{\sim}{\theta}}_{(i)}$.

According to Atkinson (1985) "since the reason for using the fully iterated estimate $\hat{\underset{\sim}{\theta}}_i$, rather than the one-step estimate $\underset{\sim}{\theta}^1_{(i)}$, is that the contours are sufficiently non-ellipsoidal to render the one-step estimate a poor approximation, it seems unwise subsequently to base a diagnostic measure on the ellipsoidal approximation." Beale (1960) pointed out that non-ellipsoidal contours for confidence regions can lead to badly biased estimators.

Cook and Weisberg (1982) however, named three alternative norms for $\hat{\underset{\sim}{\theta}} - \hat{\underset{\sim}{\theta}}^1_{(i)}$ or $\hat{\underset{\sim}{\theta}} - \underset{\sim}{\theta}_{(i)}$ that are less dependent on the shape of the contours of F, but these measures are considerably more computationally intensive. Two of these are similar to 3.2.1.4, i.e.

$$FD^1_i = \frac{1}{ps^2} \sum_{j=1}^{n} (f(\underset{\sim}{x}_j,\hat{\underset{\sim}{\theta}}) - f(\underset{\sim}{x}_j,\hat{\underset{\sim}{\theta}}^1_{(i)})^2 \qquad 3.4.8$$

or

$$FD_i = \frac{1}{ps^2} \sum_{j=1}^{n} (f(\underline{x}_j, \hat{\underline{\theta}}) - f(\underline{x}_j, \underline{\theta}_{(i)}))^2 \qquad 3.4.9$$

These measures are such that 3.2.1.4 and $FD_i^1$ are proportional when $f(\underline{x}_j, \underline{\theta})$ is linear in $\underline{\theta}$. This may however be far from true if linearity in $\underline{\theta}$ is not satisfied. If there is no controversy about parameterization, $\hat{\underline{\theta}}_{(i)}$ may be used in stead of $\hat{\underline{\theta}}^1_{(i)}$, because $FD_i$ is invariant under parameterization. If decisions about parameterization must still be taken it may be worth the trouble to use $FD_i^1$, i.e. use first approximations as decision making tool about the parameters.

Cook and Weisberg also mentioned the use of the log likelihood displacement in different forms (Cook and Weisberg, 1982 p. 188)), but make also the suggestion that they will be useful only if n, the number of observations is small.

Gonin and Money (1989) make the following suggestion for the $L_p$-norm solution where p, the exponent of the norm

$$F = \Sigma |(f(\underline{x}, \underline{\theta}) - f(\underline{x}, \hat{\underline{\theta}}))|^p \qquad 3.4.10$$

is not necessarily equal to 2. Adapt the generalized distance measure of Cook and Weisburg by substituting $\omega_p^2$ for $s^2$, the p-Jacobian matrix $J_p$ for J and variations thereof and denote it by $D_{p,i}$ in the above-mentioned measures. (See also chapter 2.1) Once again "large" values of $D_{p,i}$ would indicate that the i'th observation is influential. Note also that $\hat{\underline{e}}$ is replaced by $(\hat{\underline{y}} - \underline{f})_p$ for the p-residual vector where applicable and $h_{p,i}$ for $h_i$. The value of $\omega_p^2$ is

unknown, but can be estimated by

$$\hat{\omega}_p^2 = \frac{m_{2p-2}}{[(p-1)m_{p-2}]^2} \qquad 3.4.10$$

where p refers to the $L_p$-norm exponent and

$$m_r = \frac{1}{n}\sum_{i=1}^{n} |\hat{e}_i|^r$$

(Gonin and Money, 1989, p. 14)

Another way of looking at data is to make use of the parameter plots of Cook (1987). We shall, however not go into this here, because we want to try and find some analytical solution as indicated in chapter 4.

## 3.5 CONCLUSION

One problem in all these approaches is the amount of calculations necessary in order to recognise influential observations. Another objection is the effect of the assumption of approximate linearity. In the light of the first remark, together with the findings of Donaldson and Schnabel at the end of §3.3 we find it necessary to think of other ways to search for influence, outliers and collinearity in the non-linear situation. Although the results of §3.4 are quite useful one would like to employ a more direct approach to the problem. In chapter 4 such an approach is shown.

It is, however informative to apply the typical Cook and Dffits measures in the presence of collinearity as we shall see later with the data matrix inter alia replaced by the jacobian - hereafter called the adaptive Cook and Dffits measures.

---

# CHAPTER 4

# THE SINGULAR VALUE DECOMPOSITION AND THE DETECTION OF INFLUENTIAL OBSERVATIONS AS WELL AS VARIABLES

## 4.1 INTRODUCTION

The singular value decomposition has played an increasing role in linear regression in recent years (see Belsley *et al* 1980 and of particular importance, Belsley, 1991). The matrix mainly under concern has been the matrix of right singular vectors (i.e. the eigenvectors of X'X). Some consideration has also been given to the matrix of left singular vectors of X. In section 2 a brief summary is given of these results and is being followed by a more detailed expansion of similar investigations with respect to non-linear regression.

## 4.2 THE SINGULAR VALUE DECOMPOSITION IN LINEAR REGRESSION

Consider the standard linear regression model

$$\underline{Y} = X\underline{\beta} + \underline{\varepsilon} \qquad 4.2.1$$

where $\underline{Y}$ , X, $\underline{\beta}$ and $\underline{\varepsilon}$ have the following dimensions respectively: n×1, n×m, m×1 and n×1. The columns of X are standardized and so the vector $\underline{Y}$ is assumed to have a mean zero and a variance of one. Assume further that $\text{cov}(\underline{\varepsilon}) = \sigma^2 I$. The singular value decomposition of X, (i.e. SVD(X)) is as follows

$$X = PD_{\alpha}Q' \qquad 4.2.2$$

where P is the n×m matrix of left singular vectors, Q is the m×m matrix of right singular vectors and $D_{\alpha}$ is a diagonal matrix with the singular values down the diagonal in decreasing order. It is useful to note that $P'P = Q'Q = I_m$. A point of importance to note, is that although $QQ' = I_m$ the same cannot be said of $PP'$ in the sense that $PP' \neq I_n$. It is also note worthy that the calculation of the SVD is more stable than that of the eigenstructure (Belsley *et al* 1980).

The hat matrix

$$\begin{aligned} H &= X(X'X)^{-1}X' \\ &= PD_{\alpha}Q'(QD_{\alpha}P'PD_{\alpha}Q')^{-1}\ QD_{\alpha}P' \\ &= PD_{\alpha}Q'(QD_{\alpha}^{-2}Q')QD_{\alpha}P' \\ &= PP' \end{aligned} \qquad 4.2.3$$

is idempotent and so is $I - PP'$.

Underhill (1988) showed that the following are valid for the given situation:

$$\begin{aligned} \hat{\beta} &= (X'X)^{-1}X'\underline{Y} \\ &= (QD_{\alpha}^{2}Q')^{-1}QD_{\alpha}P'\underline{Y} \\ &= QD_{\alpha}^{-2}Q'QD_{\alpha}P'\underline{Y} \\ &= QD_{\alpha}^{-1}P'Y \end{aligned} \qquad 4.2.4$$

$$\begin{aligned} \underline{e} &= \underline{Y} - X\hat{\beta} \\ &= \underline{Y} - (PD_{\alpha}Q')(QD_{\alpha}^{-1}P'\underline{Y}) \\ &= (I - PP')\underline{Y} \end{aligned} \qquad 4.2.5$$

$$\hat{\sigma}^2 = \frac{1}{n-m}\,\underline{e}'\underline{e}$$
$$= \frac{1}{n-m}\,\hat{\underline{Y}}'(I - PP')'(I - PP')\hat{\underline{Y}}$$
$$= \frac{1}{n-m}\,\hat{\underline{Y}}'(I - PP')\hat{\underline{Y}} \qquad 4.2.6$$

$$\text{var}(\hat{\underline{\beta}}) = (X'X)^{-1}\sigma^2$$
$$= QD_a^{-2}Q'\sigma^2 \; . \qquad 4.2.7$$

The rows of $PD_a$ and $QD_a$ provide coordinates for plotting each observation (i.e. row of X) and each variable (i.e. column of X) respectively. Plotting the observations and variables simultaneously on the same system of axis is now possible.

A further use of $PD_a$ and $QD_a$ is the decomposition of the contributions of the observations and/or the variables to the total variance. This results in the possibility of the detection of influential observations in X as well as possible collinearities. The detection of outliers is also being made possible using these decompositions.



## 4.3 THE SINGULAR VALUE DECOMPOSITION OF THE JACOBIAN

Consider the model $y = f(\underline{x},\underline{\theta}) + \varepsilon$ with

$$S(\underline{\theta}) = \Sigma[y_i - f_i(\underline{x},\hat{\underline{\theta}})]^2 \qquad 4.3.1$$

where $i=1,\ldots,n$ and $\underline{\theta} = (\theta_1,\theta_2,\ldots,\theta_p)$.

Let J have the usual meaning, i.e.

$$\hat{J} = \left[\frac{\partial f_i(\underline{x},\underline{\theta})}{\partial \theta_j}\right]_{\underline{\theta}_j=\underline{\hat{\theta}}_j} \qquad 4.3.2$$

where the solution of $\underline{\theta}$ is to be found iteratively, i.e. if $\hat{J}/_{\underline{\theta}=\underline{\theta}_i} = J_i$ then using the Taylor series expansion and $\underline{\theta}_i$ as the $i^{th}$ iterated estimate of $\underline{\theta}$:

$$f(\underline{x},\underline{\theta}_{i+1}) = f(\underline{x},\underline{\theta}_i) + J_i'(\underline{\theta}_{i+1} - \underline{\theta}_i) + \ldots$$
$$f(\underline{x},\underline{\theta}_{i+1}) - f(\underline{x},\underline{\theta}_i) \doteq J_i'(\underline{\theta}_{i+1} - \underline{\theta}_i) + \underline{\varepsilon}$$

or

$$\underline{e}_i = J_i'(\underline{\theta}_{i+1} - \underline{\theta}_i) + \underline{\varepsilon}$$

so that the least squares estimate of $\underline{\theta}_{i+1} - \underline{\theta}_i$ results in the iteration

$$\underline{\theta}_{i+1} - \underline{\theta}_i = (J_i'J_i)^{-1}J_i'\underline{e}_i$$

i.e.

$$\underline{\hat{\theta}}_{i+1} = \underline{\hat{\theta}}_i + (J_i'J_i)^{-1}J_i'(\underline{y} - \underline{f}(\underline{x},\underline{\hat{\theta}}_i)) \qquad 4.3.3$$

or by means of the more complete $L_p$ model using the extended version of Gonin and Money (1989):

$$S(\underline{\theta}) = \Sigma[y_i - f_i(\underline{x},\underline{\theta})]^q$$

where q is not necessarily equal to 2 (see chapter 2.1).

If an approximate linearized model $\underline{y} = J\underline{\theta} + \underline{e}$ is assumed, then

$$\begin{aligned}\hat{\underline{\theta}} &= (J'J)^{-1}J'\underline{y} \\ &= (J'J)^{-1}J'(J\underline{\theta} + \underline{e}) \\ &= (J'J)^{-1}J'\underline{w}\end{aligned} \qquad 4.3.4$$

i.e. $\underline{w} = J\underline{\theta} + \underline{e}$ and $\underline{e} = \underline{y} - \underline{f}(\underline{x},\hat{\underline{\theta}})$. This means that J plays a similar role in determining $\hat{\underline{\theta}}$ that X plays in the determination of $\hat{\underline{\beta}}$, i.e. $\hat{\underline{\beta}} = (X'X)^{-1}X'\underline{Y}$.

The usual variance estimate will be

$$\begin{aligned}\hat{var}(\hat{\underline{\theta}}) &= (\hat{J}'\hat{J})^{-1}\hat{\sigma}^2 \\ &= (\hat{J}'\hat{J})^{-1}\frac{\Sigma e_i^2}{n-p}\ .\end{aligned} \qquad 4.3.5$$

The form of the previous equations illustrates how the behaviour of the non-linear regression parameters and variance estimators is linked to the stability of a matrix inverse. In the case of the linear regression parameter estimation the stability of the inverse of X'X plays an important role, but in non-linear regression parameter estimation it is the inverse of J'J which may be the source of instability in any estimation procedure.

This correspondence between X and J or X'X and J'J carries over to the construction of diagnostic measures in the non-linear regression case. Let us thus consider the matrix J in more detail.

Assuming $J(\underline{\theta})$ has been determined, the columns of J may be standardized in order to have means zero and variance one (see Marquardt, 1970).

The singular value decomposition of J (i.e. SVD(J)) therefore is

$$J = UD_{\alpha}V' \qquad 4.3.6$$

where U: n×p is the matrix of left singular vectors, V: p×p is the matrix of right singular vectors and $D_{\alpha}$ is the diagonal matrix with the singular values down the diagonal in decreasing order. Once again as in the previous section we note that $U'U = V'V = I_p$ , $VV' = I_p$ but $UU' \neq I_n$.

The quantities most commonly estimated in regression can now be expressed in terms of U, $D_{\alpha}$ and V.

The regression parameters:

$$\begin{aligned}
\underline{\hat{\theta}} &= (J'J)^{-1}J'\underline{w} \\
&= (J'J)^{-1}J'(J\underline{\theta} + \underline{e}) \qquad (\underline{\theta}_i \text{ against } \underline{\theta}_{i+1}) \\
&= (VD_{\alpha}U'UD_{\alpha}V')^{-1}(VD_{\alpha}U')\underline{w} \\
&= (VD_{\alpha}^{-2}V')(VD_{\alpha}U')\underline{w} \\
&= VD_{\alpha}^{-1}U'\underline{w} \qquad 4.3.7
\end{aligned}$$

or

$$\hat{\underline{\theta}} = (VD_a U' UD_a V')^{-1}(VD_a U')(UD_a V'\underline{\theta} + \underline{e})$$

$$= (VD_a^{-2} V')(VD_a U')(UD_a V'\underline{\theta} + \underline{e})$$

$$= (VD_a^{-2} V')(VD_a^2 V'\underline{\theta} + VD_a U'\underline{e})$$

$$= \underline{\theta} + VD_a^{-1} U'\underline{e} \; . \qquad 4.3.8$$

## 4.4 A PRINCIPAL COMPONENTS DECOMPOSITION OF J AND THE DETECTION OF OUTLIERS AND INFLUENCE - OBSERVATIONAL AND/OR VARIABLE

At this stage we use a principal component decomposition of the components of J together with a graphical display.

Let $J = UD_a V'$ be the SVD of the central (but not necessarily standardized) matrix J with

$$F = UD_a \qquad 4.4.1$$

$$G = VD_a \; . \qquad 4.4.2$$

The rows of F: n×p and of G: p×p provide coordinates for plotting each row of J (i.e. for each observation) and each column of J (i.e. for each variable). Note that as $FG' \neq J$ we have no biplot here. Note further that "observations" and "variables" are references to "gradiential" observations and variables because we find in non-linear regression that J consists per definition of non-linearity not merely of observations or variables, but also some form(s) of the unknown

parameters.

Further we form from $J = UD_a V'$

$$JV = UD_a = F \qquad 4.4.3$$

so that

$$\begin{aligned} JGD_a^{-1} &= JVD_aD_a^{-1} \\ &= JV \\ &= F \end{aligned} \qquad 4.4.4$$

In a similar way

$$\begin{aligned} J'FD_a^{-1} &= (VD_aU')FD_a^{-1} \\ &= GU'U \\ &= G \end{aligned} \qquad 4.4.5$$

These two results are known in linear regression, where X is the input matrix, as the transition formula and provide justification for simultaneously plotting the observations and variables on the same set of axes. In the non-linear situation we apply it in a similar way.

In a graphical display of F and G we would like it to have the following properties compared to the linear situation:

1. In linear regression distances between pairs of points in the display i.e.

between observations are Euclidean. Similarly the distances between pairs of position vectors ("point" vectors) based on $J = UD_a V'$ and $F = UD_a$ in a non-linear situation, i.e.

$$\|\underline{f}_i - \underline{f}_j\|^2 = (\underline{j}_i - \underline{j}_j)'(\underline{j}_i - \underline{j}_j) \qquad 4.4.6$$

should reflect distances proportional to distances between pairs of points in an Euclidean way. Note that $\underline{f}_j$ and $\underline{j}_j$ refer to the $j^{th}$ rows of F and J respectively. If we thus consider J as a matrix of observations where the observations are in fact the observed gradients where $\underline{\theta}$ is given or estimated, we can speak of "distances" between observed gradients. This incidentally may lead us to the second derivatives of the original equations or for that matter the rate of change in the gradiential observations for changing $\underline{\theta}$.

2. In a similar way with respect to the variables we find in linear regression systems distances between pairs of column points in an Euclidean way. Now we find the distances between pairs of position vectors (or "point" vectors) based on $J = UD_a V'$ and $G = VD_a$. So in the non-linear system

$$\|\underline{g}_i - \underline{g}_j\|^2 = (\underline{j}_i - \underline{j}_j)'(\underline{j}_i - \underline{j}_j) \; . \qquad 4.4.7$$

These distances should reflect distances (differences) proportional to distances (differences) between pairs of gradiential variables in an Euclidean way. In this case $\underline{g}_i$ refers to the $i^{th}$ row of G and $\underline{j}_i$ to the $i^{th}$ column of J.

$\phi$ being the angle between the vectors $g_i$ and $g_j$ , i.e. $S_{ij}$ is the covariance between the $i^{th}$ and $j^{th}$ gradiential variables and it follows that the correlation between these gradiential variables i and j is given by

$$r_{ij} = \frac{S_{ij}}{\sqrt{S_{ii}S_{jj}}} \qquad (S_{ii} = S_i^2,\ S_{jj} = S_j^2)$$

$$= \frac{g_i' g_j}{\|g_i\|\|g_j\|}$$

$$= \cos\phi_{ij} \ , \qquad 4.4.11$$

i.e. the cosine of the angle between the vectors $g_i$ and $g_j$ is the correlation between the gradiential variables i and j.

It follows further that the norm of J (i.e. of the set of gradiential observations) is

$$\|J\|^2 = \sum_{i=1}^{n}\sum_{j=1}^{p} j_{ij}^2 = tr(J'J) \qquad 4.4.12$$

$$= \sum_{j=1}^{p} S_j^2$$

$$= \sum_{k=1}^{p} a_k^2$$

$$= \text{total variance in J.} \qquad 4.4.13$$

In this expression $S_j^2$ is the variance of gradiential variable j and $a_k$ is the $k^{th}$ singular value of J.

The following should also be noted:

$$FF' = UD_aD_aU'$$

$$= UD_aV'VD_aU'$$

$$= JJ' \qquad 4.4.14$$

and

$$GG' = VD_aD_aV'$$

$$= VD_aU'UD_aV'$$

$$= J'J \qquad 4.4.15$$

so that

$$tr(FF') = tr(GG') \qquad 4.4.16$$

$$= \|J\|^2$$

$$= \text{total variance of the observations in J.} \qquad 4.4.17$$

(Note that J is centered, but not necessarily standardized).

The total variance inherent in J can therefore be decomposed in two ways.

F enables us to decompose the contributions of the gradiential observations to the total variance and G enables us to decompose the contributions of the gradiential variables to the total variance.

The total variance then = $tr(FF')$

$$= \sum_{i=1}^{n}\left[\sum_{k=1}^{p} f_{ik}^2\right]$$

$$= \sum_{k=1}^{p}\left[\sum_{i=1}^{n} f_{ik}^2\right] . \qquad 4.4.18$$

The term $\sum_{k=1}^{p} f_{ik}^2$ is being interpreted as the contribution of *gradiential observation* i to the total variance. This can be further decomposed into the contribution of each of the p principal axes to the variance contributed by gradiential observation i.

According to the previous paragraph we can interpret

$$\frac{f_{ik}^2}{\sum_{k=1}^{p} f_{ik}^2} \quad , \quad k=1,\ldots,p \qquad 4.4.19$$

as the proportion of the variance due to the *gradiential observation* i that is explained by the $k^{th}$ principal axis.

In a similar way of reasoning $\sum_{i=1}^{n} f_{ik}^2$ can be interpreted as the contribution *of the $k^{th}$ principal axis* to the total variance. It is also noteworthy to see that $\sum_{i=1}^{n} f_{ik}^2 = a_k^2$ .

Now $\sum_{i=1}^{n} f_{ik}^2$ may be further decomposed into the contribution of each gradiential observation to the variance of the $k^{th}$ principal axis, i.e.

$$\frac{f_{ik}^2}{\sum_{i=1}^{n} f_{ik}^2} \quad , \quad i=1,\ldots,n \qquad 4.4.20$$

as the proportion of the variance due to the $k^{th}$ principal axis that is explained by the $i^{th}$ gradiential observation.

As we turn our attention to the gradiential variables we can write

$$\text{Total variance} = tr(GG') \qquad 4.4.21$$

$$= \sum_{j=1}^{p}\left[\sum_{k=1}^{p} g_{jk}^2\right]$$

$$= \sum_{k=1}^{p}\left[\sum_{j=1}^{p} g_{jk}^2\right] \qquad 4.4.22$$

with the interpretation of

$$\frac{g_{jk}^2}{\sum_{k=1}^{p} g_{jk}^2}, \quad k=1,\ldots,p \qquad 4.4.23$$

as the proportion of the variance due to the gradiential variable j that is explained by the $k^{th}$ principal axis. We further interpret

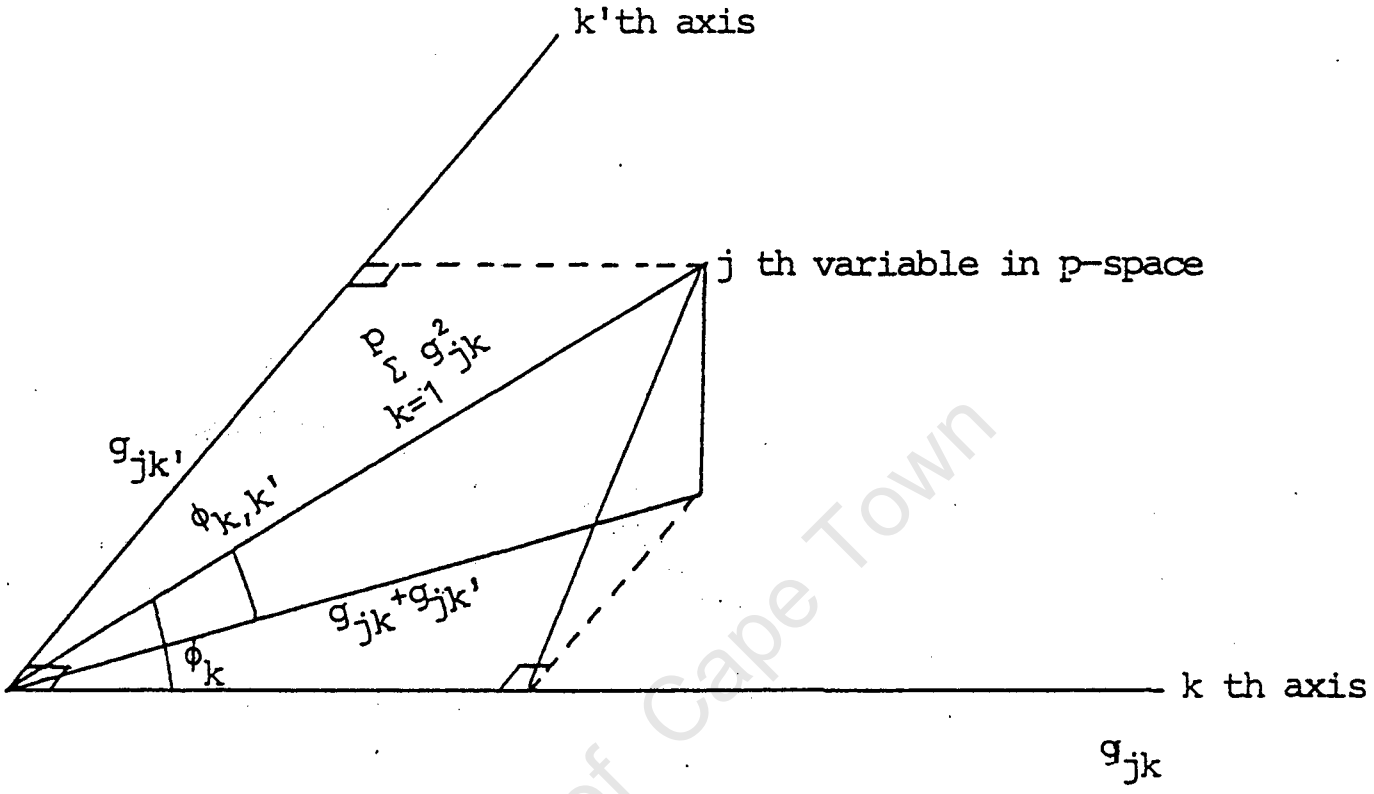
$$\frac{g_{jk}^2}{\sum_{j=1}^{p} g_{jk}^2}, \quad j=1,\ldots,p \qquad 4.4.24$$

as the proportion of the variance due to the $k^{th}$ principal axis that is explained by the gradiential variable j.

Furthermore we can see that

$$\frac{f_{ik}^2}{\sum_{k=1}^{p} f_{ik}^2} \text{ and } \frac{g_{jk}^2}{\sum_{k=1}^{p} g_{jk}^2} \qquad 4.4.25$$

are the squared cosines of the angles that the $k^{th}$ principal axis makes with the $i^{th}$ gradiential observation and $j^{th}$ gradiential variable respectively as we can see diagrammatically:






It is also useful to remember that

$$\cos^2\phi_{k,k'} = \frac{g_{jk}^2 + g_{jk'}^2}{\sum_{k=1}^{p} g_{jk}^2} \qquad 4.4.26$$

is the squared cosine of the angle between the plane defined by the $k^{th}$ principal axis and gradiential variable j. This result can now be extended to the hyperplane defined by any $m \leq p$ principal axes.

The decomposition of J thus enables us to detect

(a) influential 'observations' in J

(b) collinearities in J.

Outliers in the jacobian data matrix can be determined by considering large values of

$$\frac{f_{ik}^2}{\sum_{k=1}^{p} f_{ik}^2} \quad \text{and} \quad \frac{f_{ik}^2}{\sum_{i=1}^{n} f_{ik}^2} \qquad 4.4.27$$

for particular values of k.

A large value for the former of these two ratios implies that the $i^{th}$ gradiential observation is explained almost entirely by the $k^{th}$ principal axis. Similarly, a large value for the latter of the two ratios implies that the $k^{th}$ axis is to a large extent determined or dominated by the $i^{th}$ gradiential observation.

It is now possible to carry these indications of influence and outliers over to the original observations and variables (in terms of $\underline{w} = J\underline{\hat{\theta}} + \underline{e}$) as applicable. Note that these measures are indications in the independent variable matrix only. To determine whether there are really outliers and influence as indicated in terms of potentials should be determined as indicated in section 4.7.

If we refer to

$$\|J\|^2 = \sum_{k=1}^{p} a_k^2 \quad = \text{total variance of } J \ , \qquad 4.4.28$$

we see that collinearities will reveal themselves by having close to zero values of $a_k$. The gradiential variables $\ell$ involved in the collinearity will be those that have large values for

$$\frac{g_{\ell k}^2}{\sum_{\ell=1}^{p} g_{\ell k}^2} \qquad 4.4.29$$

because

from $JGD_a^{-1} = F$ it follows that $JG = FD_a = UD_aD_a$

and as $G = VD_a$ ; $F = UD_a$ it follows that

$$Jg_k = Jv_k a_k = u_k a_k^2$$

i.e. $Jv_k = u_k a_k$ .

But as $a_k \doteq 0$ it follows that $Jv_k \doteq 0$ ,

$$\text{i.e. } \sum_{\ell=1}^{p} j_{i\ell}\, v_{k\ell} \doteq 0 \qquad i=1,\ldots,n \ . \qquad 4.4.30$$

There is thus a linear combination of the columns of J which is near zero, implying a collinearity. The collinearity involves those columns for which $v_{k\ell}$

and hence $g_{k\ell}$ is relatively large. Redundant variables may now be eliminated where a redundant gradiential variable is being recognized by large

$$\frac{g_{\ell k}^2}{\sum_{\ell=1}^{p} g_{\ell k}^2} \qquad \ell=1,\ldots,p. \qquad 4.4.31$$

This presence of collinearity in the gradiential variables, i.e. in $w = J\underline{\hat{\theta}} + \underline{e}$ can now be drawn through in a careful consideration with respect to the original variables.

It is worth noting that the serious influential observations and outliers are revealed by the first few principal axes, and the collinearities by the last few principal axes. This result means that the detection of influential data - in particular outliers, and collinearities can proceed simultaneously.

As a result of this outliers can be removed if applicable and redundant variables eliminated by removing them or creating a dummy variable and the analysis can then be repeated if these actions have revealed any further collinearities.

It is obvious that when there is a close relationship between variables and parameters, e.g. the model $Y = \beta_0 + x_1^{\beta_1} x_2^{\beta_2} + \varepsilon$, it is easy to find the variable more or less corresponding to a bad behaving gradiential variable. In case there is no such correspondence the detection of outliers becomes very difficult indeed and one may have to look at the adapted Cook, Dffits measures and other more number crunching prone, i.e. direct methods. The detection of influential observations, however, is not affected by a similar non-correspondence.

## 4.5 DEALING WITH INFLUENTIAL OBSERVATIONS

In section 4.4 the proportional contributions of each row and column are determined by means of the singular value decomposition of the jacobian matrix J, such that

$$J = UD_{a}V' \qquad 4.5.1$$

and the corresponding principal components regression analysis in terms of

$$F = UD_{a} \text{ and } G = VD_{a} \qquad 4.5.2$$

These contributions then may indicate influential observations and in special circumstances as described above they can be indicative of outliers. A further result is that collinearity is being spotted and a decision can be made about the principal axes' contributions. In this and the following two sections some proposals for future research are made with respect to outliers, influence and robust estimation.

In the case of influential observations identified or collinearities which may be present there may be several approaches to the problem. One way of dealing with the aforementioned situation(s) is to replace the corresponding row (influential observation) by the average contribution of that row in the original data matrix. (Replacing a large row contribution by the proportional average of the columns may reduce the lengths of the corresponding principal axis, but the interpretation may not be so clear.) Call the new (adapted) data matrix, $X^{*}$, then it follows that

$$J^* = U^* D_a^* V^{*\prime} \qquad 4.5.3$$

where $J^*$ is the jacobian based on $X^*$. Use $J^*$ to estimate $\underline{\theta}$, i.e.

$$\hat{\underline{\theta}}^* = (J^{*\prime} J^*)^{-1} J^{*\prime} \underline{w}^* \qquad 4.5.4$$

where $\underline{w}^* = J^* \underline{\theta}^* + \underline{e}^*$ and $\underline{e}^* = \underline{y} - \underline{f}^*(\underline{x},\underline{\theta})$ with $\underline{f}^*(\underline{x},\underline{\theta})$ adapted according to the previous paragraph.

Effectively we have used a TRIMMED or ROBUST estimator.

## 4.6 DEALING WITH INFLUENTIAL OBSERVATIONS TOGETHER WITH COLLINEARITY

When one or more influential observations are detected and simultaneously one observes collinearity one can deal with both problems in one step.

Construct a matrix $X^*$ as described in the previous section and bring in a ridge value according to the Levenburg-Marquardt technique. Then

$$J_L^* = U_L^* (D_a + \lambda I)^* V_L^{*\prime}, \quad U_L^* = U^* \text{ and } V_L^* = V^* \qquad 4.6.1$$

or use the r columns and rows associated with the r largest roots, i.e.

$$J_{[r]}^* = U_{[r]}^* D_{a[r]}^* V_{[r]}^{*\prime} \qquad 4.6.2$$

in order to determine $\hat{\theta}_L^*$ or $\hat{\theta}_{[r]}^*$. One can take this even further by combining the last two robust estimator carriers, i.e.

$$J^{*}_{L[r]} = U^{*}_{L[r]} \, (D_a + \lambda I)^{*}_{[r]} V^{*,}_{L[r]} \qquad 4.6.3$$

to determine $\hat{\theta}^{*}_{L[r]}$, i.e.

$$\hat{\theta}^{*}_{L[r]} = (J^{*,}_{L[r]} J^{*}_{L[r]})^{-1} J^{*,}_{L[r]} w^{*}_{L[r]} \qquad 4.6.4$$

where $w^{*}_{L[r]} = J^{*}_{L[r]} \theta^{*}_{L[r]} + \underline{e}^{*}_{L[r]}$ and $\underline{e}^{*}_{L[r]} = \underline{y} - f^{*}_{L[r]}(\underline{x}, \underline{\theta})$ with $f^{*}_{L[r]}(\underline{x}, \underline{\theta})$ adapted in accordance with the interpretation of *, L and [r].

This last robust estimator is a **TRIMMED, LEVENBURG-MARQUARDT, DIMINISHED RANK ROBUST ESTIMATOR.**

## 4.7 OUTLIERS

It was shown in section 3.4 that the (i,i)th element in $J(J'J)^{-1}J'$ can serve as an indication of leverage and that a relatively large value of the (i,i)th element is an indication of a possible influential observations and/or an outlier in terms of the data matrix.

With respect to the observed values, $y_i$, of the model one can use the following measures based on $[J(J'J)^{-1}J']_{ii}$ to determine whether outliers are in actual fact present by testing only for those observations indicated as possibles by the methods described in the previous sections, thereby reducing a heavy load of unecessary calculations.

Define

$$h_{ii} = [J(J'J)^{-1}J']_{ii} = [H]_{ii}$$

$$h^{*}_{ii} = [J^{*}(J^{*'}J^{*})^{-1}J^{*'}]_{ii} = [H^{*}]_{ii}$$

$$h_{ii(L)} = [J_L(J_L'J_L)^{-1}J_L']_{ii} = [H_L]_{ii}$$

$$h_{ii[r]} = [J_{[r]}(J_{[r]}'J_{[r]})^{-1}J_{[r]}']_{ii} = [H_{[r]}]_{ii}$$

$$h^{*}_{ii(L)} = [J^{*}_L(J^{*'}_L J^{*}_L)^{-1}J^{*'}_L]_{ii} = [H^{*}_L]_{ii}$$

$$h^{*}_{ii[r]} = [J^{*}_{[r]}(J^{*'}_{[r]}J^{*}_{[r]})^{-1}J^{*'}_{[r]}]_{ii} = [H^{*}_{[r]}]_{ii}$$

$$h_{ii(L)[r]} = [J_{L[r]}(J_{L[r]}'J_{L[r]})^{-1}J_{L[r]}']_{ii} = [H_{L[r]}]_{ii}$$

$$h^{*}_{ii(L)[r]} = [J^{*}_{L[r]}(J^{*'}_{L[r]}J^{*}_{L[r]})^{-1}J^{*'}_{L[r]}]_{ii} = [H^{*}_{L[r]}]_{ii} \qquad 4.7.1$$

Then we can write the measures as follows for each of the versions mentioned above

$$t^{\cdot}_{i..} = \frac{e^{\cdot}_{i..}}{s\sqrt{1 - h^{\cdot}_{ii..}}} \qquad 4.7.2$$

and similarly, keeping in mind that no leave out methods are used here as customary in the following measures which may correspond to almost similar measures in linear regression:

$$DFFIT^{\cdot}_{i..} = \frac{h^{\cdot}_{i..}\ e^{\cdot}_{i..}}{1 - h^{\cdot}_{i..}} \qquad 4.7.3$$

$$DFFITS^{2\cdot}_{ii..} = \frac{h^{\cdot}_{ii..}}{(1 - h^{\cdot}_{i..})^2} \cdot \frac{e^{\cdot 2}_{i..}}{s^2} \qquad 4.7.4$$

$$COOK^{\cdot}_{..} = DFFITS^{2\cdot}_{..}/p \qquad 4.7.5$$

In the measures mentioned here $s^2$ may be replaced by $s^{2\cdot}_{(i)..}$ which will be the estimated variance based on a matrix where the ith observation is being changed in $X^*$ and/or the other possible conversions. Comparisons between measures, however may then be inappropriate as the scaling factor will then be dependent on the observation which is being changed.

## 4.8 AN EXAMPLE WITH SOME RESULTS

### 4.8.1 The data

In order to illustrate the use of at least some of the measures and indicators of the previous sections data was created using the technique as described by Thiart et al (1991) based on a simulation study in the unpublished Ph.D. thesis of Chalton (1990). Whereas Chalton used 5 parameters, we use a five-variable six parameters model which includes a non-zero constant called $\beta_6$. A sample size of 30 was chosen. The variables were simulated as follows:

For j = 1,2,3 and i = 1,...,30

$$X_{ij} = (1 - a_1^2)^{\frac{1}{2}} Z_{ij} + a_1 X_{i6}$$

and $j = 4,5$ and $i = 1,\ldots,30$

$$X_{ij} = (1 - a_2^2)^{\frac{1}{2}} Z_{ij} + a_2 Z_{i6}$$

where $Z_{ij} \sim n(0,1)$ variates generated by SAS-function **RANNOR**. The seeds are not available as the **RANNOR**-function derives the seeds from the real time clock of the computer. The degree of collinearity is determined by the values of parameters $a_1$ and $a_2$, where $a_1^2$ is the theoretical correlation between any pair of the variables $X_1$, $X_2$ and $X_3$, $a_1 a_2$ is the theoretical correlation between any pair of the variables $X_1, X_2$ or $X_3$ and $X_4$ or $X_5$ and $a_2^2$ is the theoretical correlation between $X_4$ and $X_5$.

In this empirical investigation the values chosen for the parameters where $a_1^2 = a_2^2 = 0,99$ i.e. a high collinearity factor was built into the data. The values for $\underline{\theta}$ were determined by a suggestion of Newhouse and Oman (1971), namely the eigenvectors of $X'X$ according to the largest eigenvalues. The model and resulting data set are as follows:

$$Y_i = X_{i1}^{\theta_1} X_{i2}^{\theta_2} X_{i3}^{\theta_3} X_{i4}^{\theta_4} X_{i5}^{\theta_5} \theta_6 + \varepsilon_i \qquad i = 1,\ldots,30$$

where, after having increased all x-values by 2

$$X_{ij} \sim n(2,1); \quad \underline{\theta}' = (0,4474;\ 0,4473;\ 0,4481;\ 0,4470;\ 0,4463;\ 10)$$

| n | y | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ |
|---|---|---|---|---|---|---|
| 1 | 59.37928 | 1.91510 | 2.07781 | 2.33137 | 2.25610 | 2.44236 |
| 2 | 17.03769 | 1.20009 | 1.12683 | 1.11298 | 1.46529 | 1.48450 |
| 3 | 31.43451 | 1.82134 | 1.57941 | 1.84071 | 1.86991 | 1.32875 |
| 4 | 4.21973 | .46757 | .79256 | .53950 | .72713 | .52242 |
| 5 | 136.39150 | 3.33529 | 3.12541 | 3.15277 | 3.05349 | 3.39790 |
| 6 | 105.87280 | 2.84638 | 2.85153 | 2.68915 | 2.94020 | 2.94423 |
| 7 | 123.77210 | 3.08654 | 2.87469 | 2.91956 | 3.36636 | 3.09060 |
| 8 | 16.74560 | 1.05388 | 1.23509 | 1.36169 | 1.40634 | 1.32092 |
| 9 | 45.54456 | 2.12865 | 2.12626 | 2.06978 | 1.77268 | 1.75008 |
| 10 | 55.86286 | 2.23236 | 2.12059 | 2.11275 | 2.04434 | 2.24608 |
| 11 | 12.31348 | 1.25142 | .87584 | 1.16432 | 1.13806 | .82105 |
| 12 | 63.57678 | 2.33503 | 2.25782 | 2.37770 | 2.20623 | 2.25707 |
| 13 | 8.19416 | 1.19581 | .52267 | 1.07145 | 1.24602 | .91252 |
| 14 | 81.52631 | 2.51869 | 2.42961 | 2.45906 | 2.72280 | 2.54492 |
| 15 | 28.93762 | 1.69534 | 1.58217 | 1.45599 | 1.51818 | 1.75220 |
| 16 | 3.47572 | .26741 | .37465 | .38450 | .33489 | .43181 |
| 17 | 103.02420 | 2.82416 | 2.80942 | 2.86994 | 2.77163 | 2.96814 |
| 18 | 16.99822 | 1.18043 | 1.08162 | 1.51366 | 1.20798 | 1.37967 |
| 19 | 37.89568 | 1.92986 | 1.91445 | 1.79639 | 1.89782 | 1.50474 |
| 20 | 286.31040 | 4.41738 | 4.51685 | 4.44304 | 4.58753 | 4.43136 |
| 21 | 130.36960 | 3.07069 | 3.28934 | 3.22908 | 2.97070 | 3.18371 |
| 22 | 25.81647 | 1.57042 | 1.39488 | 1.64897 | 1.48323 | 1.54718 |
| 23 | 85.89640 | 2.56950 | 2.62498 | 2.66179 | 2.43721 | 2.82084 |
| 24 | 15.18494 | .96893 | 1.14209 | 1.19054 | 1.18316 | 1.19166 |
| 25 | 249.91060 | 4.14797 | 4.18853 | 4.11878 | 4.34428 | 4.35794 |
| 26 | 171.23430 | 3.67771 | 3.53004 | 3.48130 | 3.49611 | 3.62604 |
| 27 | 103.81700 | 2.73229 | 2.90525 | 2.88059 | 2.84373 | 2.86646 |
| 28 | 153.33220 | 3.50644 | 3.31758 | 3.41795 | 3.50178 | 3.19613 |
| 29 | 64.36023 | 2.21360 | 2.17496 | 2.40920 | 2.38721 | 2.35472 |
| 30 | 3.65093 | .39235 | .62900 | .85005 | .54118 | .39154 |

table 4.1

### 4.8.2 Procedures and results

The solution was determined using a FORTRAN-program code which was adapted from the code in Ratkowsky (1983). As the Linpack routines were not available they were replaced by my own routines using code from the Numerical Recipes routine package of Vetterling, Teukolsky, Press and Flannery (1985). The results were checked against Ratkowsky's version as well as the SAS-package. See appendix D for programming codes.

The first results are as given in the following table.



PARAMETER ESTIMATES AT CONVERGENCE.

| $\theta$ | ESTIMATE | SE | T |
|---|---|---|---|
| 1 | .421406E+00 | .476008E-01 | 8.85 |
| 2 | .523378E+00 | .642551E-01 | 8.15 |
| 3 | .353925E+00 | .665027E-01 | 5.32 |
| 4 | .499783E+00 | .364521E-01 | 13.71 |
| 5 | .412923E+00 | .366393E-01 | 11.27 |
| 6 | .103273E+02 | .114258E+00 | 90.39 |

table 4.2

PARAMETER VARIANCE-COVARIANCE MATRIX

| | | | | | | |
|---|---|---|---|---|---|---|
| 1 | .22658E-02 | | | | | |
| 2 | -.10564E-02 | .41287E-02 | | | | |
| 3 | -.33792E-03 | -.28979E-02 | .44226E-02 | | | |
| 4 | -.66540E-03 | .26375E-03 | -.71942E-03 | .13287E-02 | | |
| 5 | -.10213E-03 | -.71215E-03 | -.17744E-03 | -.30031E-03 | .13424E-02 | |
| 6 | -.13046E-02 | .35613E-02 | -.34366E-02 | .10128E-02 | -.82449E-03 | .13055E-01 |

table 4.3

PARAMETER CORRELATIONS

| | | | | | | |
|---|---|---|---|---|---|---|
| 1 | 1.000000 | | | | | |
| 2 | -.345390 | 1.000000 | | | | |
| 3 | -.106749 | -.678173 | 1.000000 | | | |
| 4 | -.383488 | .112608 | -.296774 | 1.000000 | | |
| 5 | -.058561 | -.302497 | -.072825 | -.224854 | 1.000000 | |
| 6 | -.239886 | .485087 | -.452275 | .243178 | -.196949 | 1.000000 |

table 4.4

| UNIT | Y | FITTED | RESIDUAL | JHAT(I,I) | DFFITS**2 |
|---|---|---|---|---|---|
| 1 | .593793E+02 | .583413E+02 | .103801E+01 | .378262E+00 | .114361E+01 |
| 2 | .170377E+02 | .175686E+02 | -.530897E+00 | .801207E-01 | .289466E-01 |
| 3 | .314345E+02 | .322265E+02 | -.791977E+00 | .204048E+00 | .219117E+00 |
| 4 | .421973E+01 | .347983E+01 | .739904E+00 | .243529E-01 | .151918E-01 |
| 5 | .136391E+03 | .135395E+03 | .996384E+00 | .428983E+00 | .141673E+01 |
| 6 | .105873E+03 | .105534E+03 | .338631E+00 | .346153E+00 | .100708E+00 |
| 7 | .123772E+03 | .123246E+03 | .526276E+00 | .400701E+00 | .335162E+00 |
| 8 | .167456E+02 | .174970E+02 | -.751356E+00 | .725430E-01 | .516409E-01 |
| 9 | .455446E+02 | .457254E+02 | -.180794E+00 | .231229E+00 | .138711E-01 |
| 10 | .558629E+02 | .558612E+02 | .168228E-02 | .890395E-01 | .329363E-06 |
| 11 | .123135E+02 | .109899E+02 | .132355E+01 | .486902E-01 | .102229E+00 |
| 12 | .635768E+02 | .638512E+02 | -.274399E+00 | .898781E-01 | .886162E-02 |
| 13 | .819416E+01 | .873318E+01 | -.539020E+00 | .151249E+00 | .661659E-01 |
| 14 | .815263E+02 | .809195E+02 | .606773E+00 | .152108E+00 | .844924E-01 |
| 15 | .289376E+02 | .290958E+02 | -.158175E+00 | .111335E+00 | .382582E-02 |
| 16 | .347572E+01 | .103391E+01 | .244181E+01 | .824383E-03 | .534027E-02 |
| 17 | .103024E+03 | .104043E+03 | -.101895E+01 | .103872E+00 | .145666E+00 |
| 18 | .169982E+02 | .167735E+02 | .224749E+00 | .103393E+00 | .704650E-02 |
| 19 | .378957E+02 | .383979E+02 | -.502232E+00 | .254653E+00 | .125411E+00 |
| 20 | .286310E+03 | .285388E+03 | .922760E+00 | .615522E+00 | .384568E+01 |
| 21 | .130370E+03 | .130056E+03 | .314056E+00 | .354018E+00 | .907596E-01 |
| 22 | .258165E+02 | .258789E+02 | -.624123E-01 | .626106E-01 | .301052E-03 |
| 23 | .858964E+02 | .862728E+02 | -.376373E+00 | .178307E+00 | .405770E-01 |
| 24 | .151849E+02 | .135884E+02 | .159658E+01 | .388159E-01 | .116164E+00 |
| 25 | .249911E+03 | .251360E+03 | -.144902E+01 | .423540E+00 | .290268E+01 |

| | | | | | |
|---|---|---|---|---|---|
| 26 | .171234E+03 | .171172E+03 | .623169E-01 | .289151E+00 | .241033E-02 |
| 27 | .103817E+03 | .104394E+03 | -.577438E+00 | .187088E+00 | .102391E+00 |
| 28 | .153332E+03 | .153283E+03 | .495911E-01 | .398720E+00 | .294181E-02 |
| 29 | .643602E+02 | .651069E+02 | -.746719E+00 | .170422E+00 | .149768E+00 |
| 30 | .365093E+01 | .257621E+01 | .107472E+01 | .103238E-01 | .132051E-01 |

RSS = 22.126730 VAR = 0.92194720

table 4.5

Observed residuals: With $s = \sqrt{0,9219472} \doteq 0,96$ no observations are seen as having extreme residuals, i.e. more than 3s, except observation no 16 with a residual of 2,44 which is 2,54 standard deviations.

Influence: Keep in mind the cut-off values of Belsley, Kuh and Welsch (1980), i.e. $2\sqrt{\frac{p}{n}}$ for DFFITS or $\frac{4p}{n}$ for DFFITS$^2$.

For this example the cut-off will then be $\frac{4(6)}{30} = 0,8$. This leads to observations 1, 5, 20 and 25 as being influential observations.

Leverage: Refer again to Belsley et al (1980). The ith observation is a leverage point when $h_{ii}$, or JHAT(i,i) exceeds $\frac{2p}{n}$, In this case $\frac{2p}{n} = \frac{2(6)}{30} = 0,4$. Leverage points are then observations no. 5, 7, 20 and 25.

Let us determine $\frac{\sum_j f_{ij}^2}{\sum\sum_{ij} f_{ij}^2}$. The following tables give the necessary detail:

| i | | | $\frac{f_{ij}^2}{\sum_j f_{ij}^2}$ | | | | $\sum_j f_{ij}^2$ | $\frac{\sum_j f_{ij}^2}{\sum\sum_{ij} f_{ij}^2}$ |
|---|---|---|---|---|---|---|---|---|
| 1 | .982 | .002 | .008 | .006 | .000 | .002 | 6231.972000 | .0038 |
| 2 | .999 | .000 | .001 | .000 | .000 | .000 | 29473.830000 | .0180 |
| 3 | .996 | .003 | .001 | .000 | .000 | .000 | 20766.320000 | .0127 |
| 4 | 1.000 | .000 | .000 | .000 | .000 | .000 | 34232.840000 | .0209 |
| 5 | .995 | .003 | .000 | .002 | .000 | .000 | 29850.990000 | .0182 |
| 6 | .990 | .000 | .004 | .001 | .001 | .005 | 4485.708000 | .0027 |
| 7 | .989 | .006 | .004 | .001 | .000 | .000 | 16508.320000 | .0101 |
| 8 | .999 | .000 | .000 | .000 | .000 | .000 | 29469.780000 | .0180 |
| 9 | .993 | .000 | .007 | .000 | .000 | .000 | 12648.320000 | .0077 |
| 10 | .997 | .001 | .000 | .001 | .000 | .000 | 7293.279000 | .0045 |
| 11 | 1.000 | .000 | .000 | .000 | .000 | .000 | 32455.650000 | .0198 |
| 12 | .996 | .000 | .003 | .000 | .000 | .001 | 3966.757000 | .0024 |
| 13 | .999 | .000 | .000 | .000 | .000 | .000 | 33166.520000 | .0202 |
| 14 | .799 | .114 | .076 | .002 | .008 | .001 | 199.869600 | .0001 |
| 15 | .999 | .000 | .000 | .000 | .000 | .000 | 22661.220000 | .0138 |
| 16 | 1.000 | .000 | .000 | .000 | .000 | .000 | 33638.760000 | .0205 |
| 17 | .993 | .007 | .000 | .000 | .000 | .001 | 3956.326000 | .0024 |
| 18 | 1.000 | .000 | .000 | .000 | .000 | .000 | 29662.660000 | .0181 |
| 19 | .996 | .002 | .002 | .000 | .000 | .001 | 17079.160000 | .0104 |
| 20 | 1.000 | .000 | .000 | .000 | .000 | .000 | 602605.400000 | .3678 |
| 21 | .994 | .004 | .002 | .001 | .000 | .000 | 23422.130000 | .0143 |
| 22 | 1.000 | .000 | .000 | .000 | .000 | .000 | 24488.190000 | .0149 |
| 23 | .219 | .768 | .002 | .000 | .002 | .009 | 111.799900 | .0001 |
| 24 | 1.000 | .000 | .000 | .000 | .000 | .000 | 31405.880000 | .0192 |
| 25 | 1.000 | .000 | .000 | .000 | .000 | .000 | 397075.800000 | .2424 |
| 26 | .999 | .000 | .000 | .001 | .000 | .000 | 93356.640000 | .0570 |
| 27 | .992 | .001 | .000 | .006 | .001 | .000 | 4033.730000 | .0025 |
| 28 | .997 | .002 | .001 | .000 | .000 | .000 | 56451.310000 | .0345 |
| 29 | .992 | .000 | .002 | .002 | .000 | .004 | 3535.054000 | .0022 |
| 30 | 1.000 | .000 | .000 | .000 | .000 | .000 | 34078.050000 | .0208 |

table 4.6

| i | $\frac{f_{ij}^2}{\sum_i f_{ij}^2}$ | | | | | |
|---|---|---|---|---|---|---|
| 1 | .003738 | .017046 | .095148 | .123814 | .019843 | .096702 |
| 2 | .017992 | .001269 | .042058 | .000914 | .003781 | .028965 |
| 3 | .012637 | .085691 | .044903 | .000344 | .002336 | .025294 |
| 4 | .020910 | .000633 | .000002 | .006563 | .048524 | .055510 |
| 5 | .018146 | .113730 | .000248 | .226983 | .008269 | .028672 |
| 6 | .002713 | .000444 | .031386 | .012788 | .093596 | .174538 |
| 7 | .009974 | .149501 | .130110 | .044156 | .060809 | .000712 |
| 8 | .017996 | .000485 | .008348 | .037398 | .001238 | .002134 |
| 9 | .007672 | .001804 | .166873 | .001002 | .009209 | .015436 |
| 10 | .004444 | .015467 | .000194 | .028312 | .005136 | .004692 |
| 11 | .019822 | .013091 | .005435 | .002802 | .036907 | .001916 |
| 12 | .002414 | .002816 | .019608 | .000622 | .009327 | .023363 |
| 13 | .020248 | .017646 | .005793 | .015933 | .081429 | .048115 |
| 14 | .000098 | .034500 | .028645 | .001113 | .065154 | .001643 |
| 15 | .013832 | .006951 | .003220 | .027892 | .000135 | .063462 |
| 16 | .020550 | .000007 | .000059 | .001258 | .104097 | .021776 |
| 17 | .002399 | .038809 | .000990 | .000468 | .013872 | .021672 |
| 18 | .018117 | .002379 | .002080 | .004143 | .030176 | .046983 |
| 19 | .010390 | .043791 | .065153 | .000193 | .032522 | .072323 |
| 20 | .368156 | .027795 | .028672 | .093502 | .082167 | .005310 |
| 21 | .014219 | .129982 | .078263 | .068148 | .034687 | .002753 |
| 22 | .014960 | .000179 | .000106 | .000929 | .007880 | .022349 |
| 23 | .000015 | .129496 | .000480 | .000053 | .008093 | .007923 |
| 24 | .019184 | .000087 | .002582 | .017036 | .008326 | .011887 |
| 25 | .242579 | .000220 | .108378 | .003587 | .036491 | .010635 |
| 26 | .056997 | .021324 | .011528 | .163297 | .002945 | .000534 |
| 27 | .002446 | .007127 | .001228 | .072323 | .085357 | .004546 |
| 28 | .034392 | .136451 | .104658 | .017380 | .014386 | .090641 |
| 29 | .002144 | .001118 | .012760 | .017300 | .017635 | .099278 |
| 30 | .020818 | .000164 | .001092 | .009748 | .075676 | .010235 |
| | 1636643.0 | 662.7 | 531.7 | 319.7 | 24.0 | 131.5 |

$\sum_i f_{ij}^2$

$$\frac{\sum_i f_{ij}^2}{\sum\sum_{ij} f_{ij}^2}$$

| | | | | | |
|---|---|---|---|---|---|
| .99898 | .00040 | .00032 | .00020 | .00001 | .00008 |

table 4.7

There is no clear cut way of choosing the observations with large variance contributions – what is "large"? We can see in table 4.6 however, that gradiential observations no. 20 and 25 make up 36.78% plus 24.24% = 61.02% of the total variance. This indicates a large potential for influence through the jacobian inherent in the data set together with the model.

Observations 20 and 25 are the two observations which occur every time as having relatively large leverage, being influential and having relatively large residuals.

When we remove observations 20 and 25 the new estimates for $\underline{\theta}$ are as follows with comparative values before deletion:

| i | simulated $\theta_i$ | 30 observations: $\hat{\theta}_i$ | 28 observations (20;25 deleted): $\hat{\theta}_i$ |
|---|---|---|---|
| 1 | 0.4474 | 0.4214 | 0.333 |
| 2 | 0.4473 | 0.5234 | 0.633 |
| 3 | 0.4481 | 0.3539 | 0.418 |
| 4 | 0.4470 | 0.4998 | 0.586 |
| 5 | 0.4463 | 0.4129 | 0.213 |
| 6 | 10.0 | 10.3 | 10.6 |

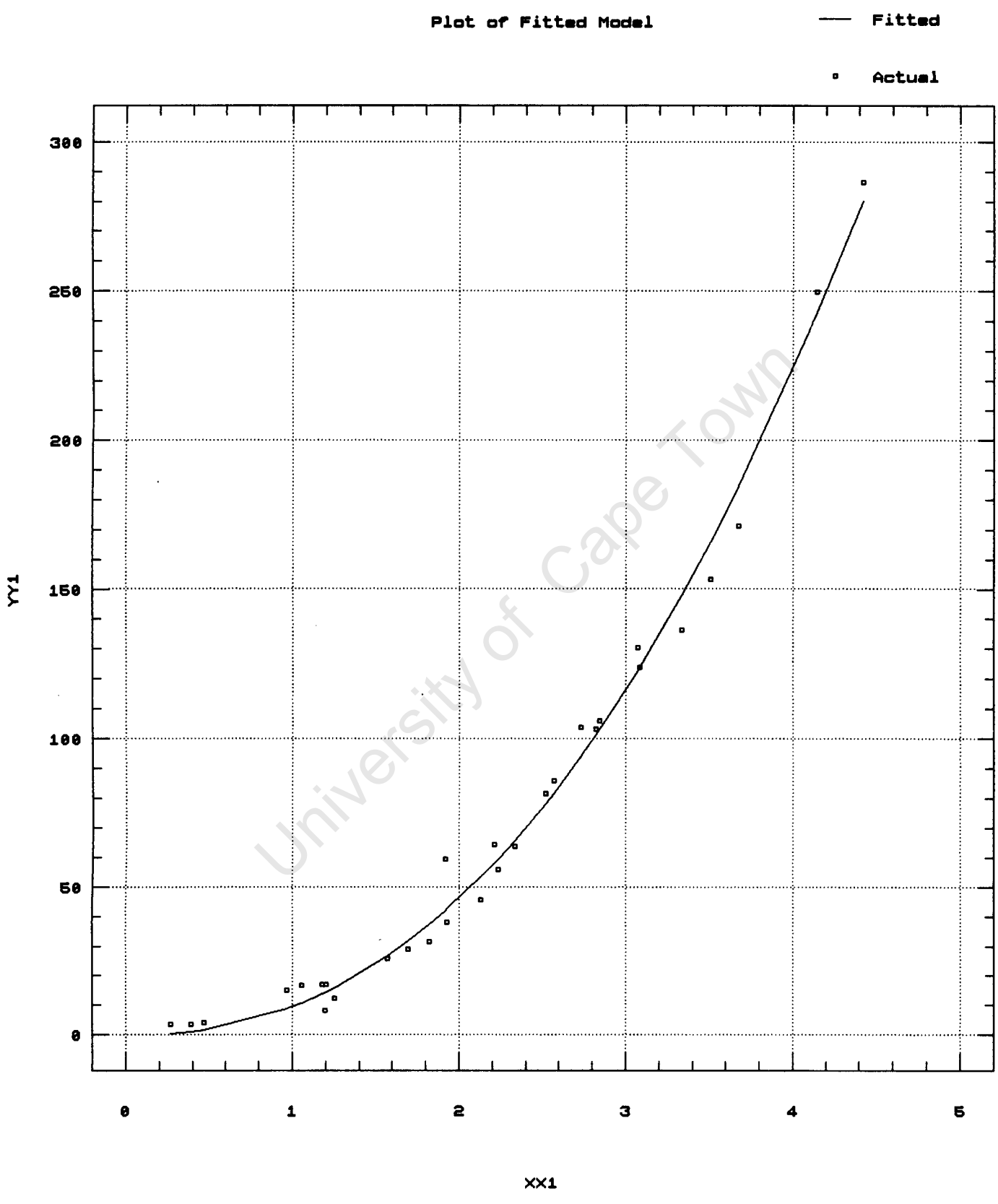
table 4.8

In table 4.8 it is clear how the removal of only these two observations influence the estimates. Here too we can see that the influence of these two observations is most important. The presence of observations 20 and 25 results in more accurate estimates for all the parameters except for $\theta_3$ where the estimate moves away from the correct value. It is of interest that the values in table 4.6 and further were determined by using the centralized gradiential data matrix. If the raw gradiential data is used the values for $\frac{\sum_j f_{ij}^2}{\sum\sum_{ij} f_{ij}^2}$ differ only marginally in the sense of correctional rounding off values. Standardization was not deemed necessary as the X-variables were all simulated with similar distributions. A comparison then also showed that the results are similar up to at least the 3rd decimal.

A further point of interest is that in table 4.7 $\frac{\sum_{i=1}^{n} f_{i1}^2}{\sum\sum_{jk} f_{jk}^2} = 0.99$ which indicates that the first principal axis declares 99% of the total variation, i.e. the model

$$\underline{Y} = \underline{X}_1 \beta_2 + \underline{\varepsilon}$$

may be appropriate for this data set, including the constant largely because of its large significant t-value. When this last model is fitted using 30 observations the mean squared error equals 50.09. Plotting the graph we find that we have a fairly good fit, see the plot on the next page:



To further substantiate the removal of variables $x_2$ to $x_5$ we look at the following table:

| | | | $g_{jk}$ | | | |
|---|---|---|---|---|---|---|
| | 569.4088 | -3.5002 | 8.3697 | -14.3582 | -0.3783 | -0.3105 |
| | 572.5391 | 5.5294 | 9.1586 | 6.4482 | 0.3927 | 7.7728 |
| | 564.0843 | 2.5704 | 7.5782 | 7.2602 | -0.3317 | -8.2727 |
| | 577.8245 | -19.5971 | -9.5992 | 3.2776 | 0.2496 | 0.9567 |
| | 575.4363 | 15.0987 | -15.1334 | -2.5228 | -0.2545 | -0.1950 |
| | 37.1223 | 1.2530 | 0.1254 | -1.4481 | 4.8454 | -1.2801 |
| $\sum_j g_{jk}^2$: | | | | | | |
| | 1 636 604 | 662 | 531.6 | 319.6 | 24.0 | 131.53 |
| | | | | | TOTAL = | 1 638 272.7 |
| $\frac{\sum_j g_{jk}^2}{\sum\sum_{jk} g_{ik}^2} \times 100\%$: | | | | | | |
| | 99.89 | 0.04 | 0.03 | 0.02 | 0.0 | 0.01 |

table 4.9

As $\frac{\sum_{i=1}^{p} g_{jk}^2}{\sum_{j=1}^{p}\sum_{k=1}^{p} g_{jk}^2} \times 100\%$ represents the percentage variance due to gradiential variable k in all the principal axes together, we see from table 4.9 that the contributions by gradiential variables $x_2$, $x_3$, $x_4$, $x_5$ and $x_6$ are 0.04%; 0.03%; 0.02%; 0% and 0.0% respectively.

The way that observations 20 and 25 stand out as being contributors to the total variance can also be observed using 3-dimensional plots of the columns of the

matrix consisting of $F = UD_a$ with $G = VD_a$ concatenated vertically. The Statgraphics graphical procedures were used. The symbols JFi refer to the ith principal axes based on the jacobian J.

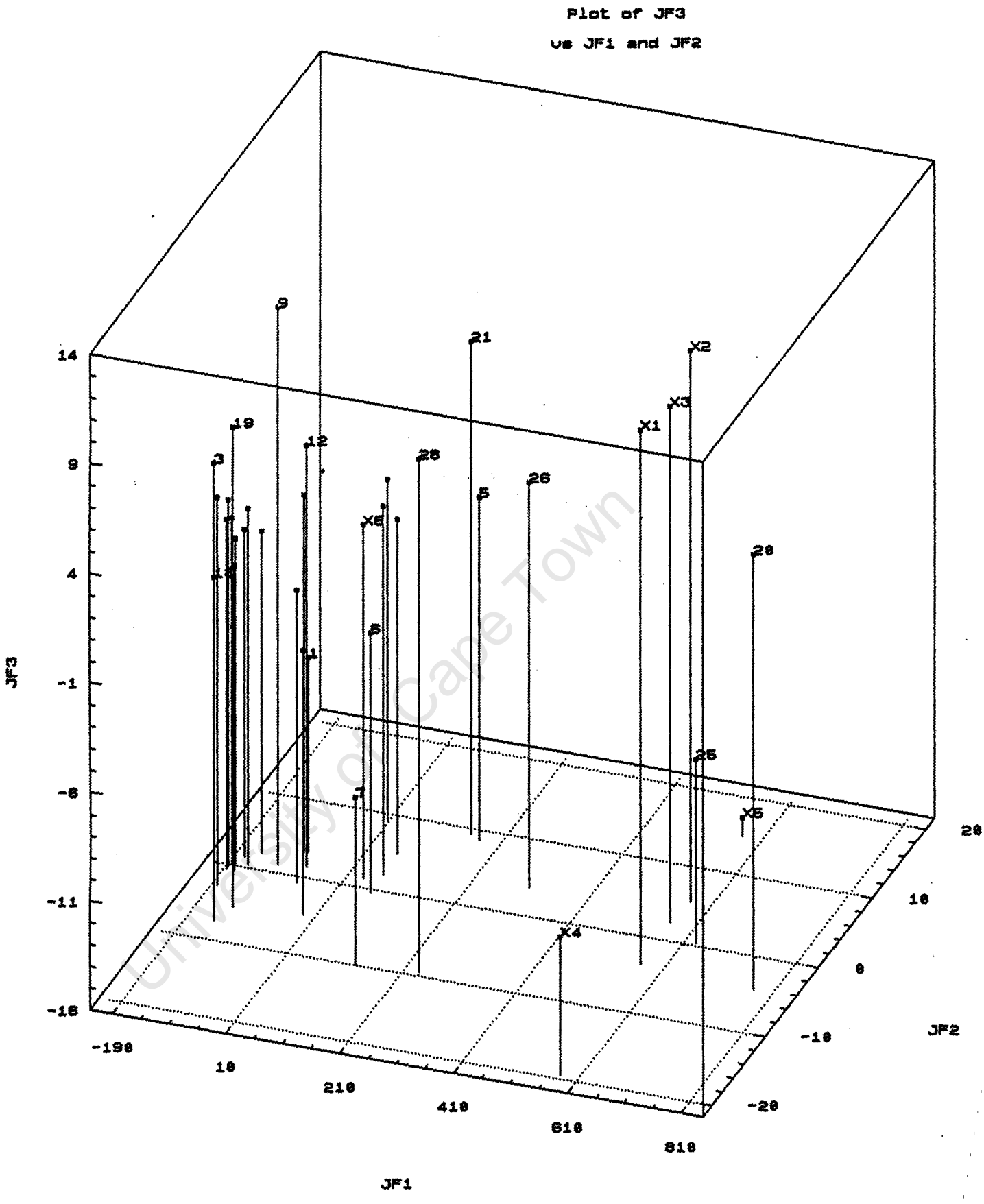
Note that the first principal axis is responsible for almost 100% of the total variance so that the sizes (scales) and therefore the importance of the second, third and fourth principal axes are over-emphasized by the graphical display. It is interesting to see how the results of the last few pages are being lifted out in these graphs as the relative positions of observations 20 and 25 are being observed largely with respect to principal axis one. Note also their positions as being right on the outside of the convex hull - in this case in fact almost a linear convex hull because of the relative importance of principal axis one. Note also how the importance of observations 20 and 25 vanishes when the axes of the graphical representations do not include the important first principal axis.

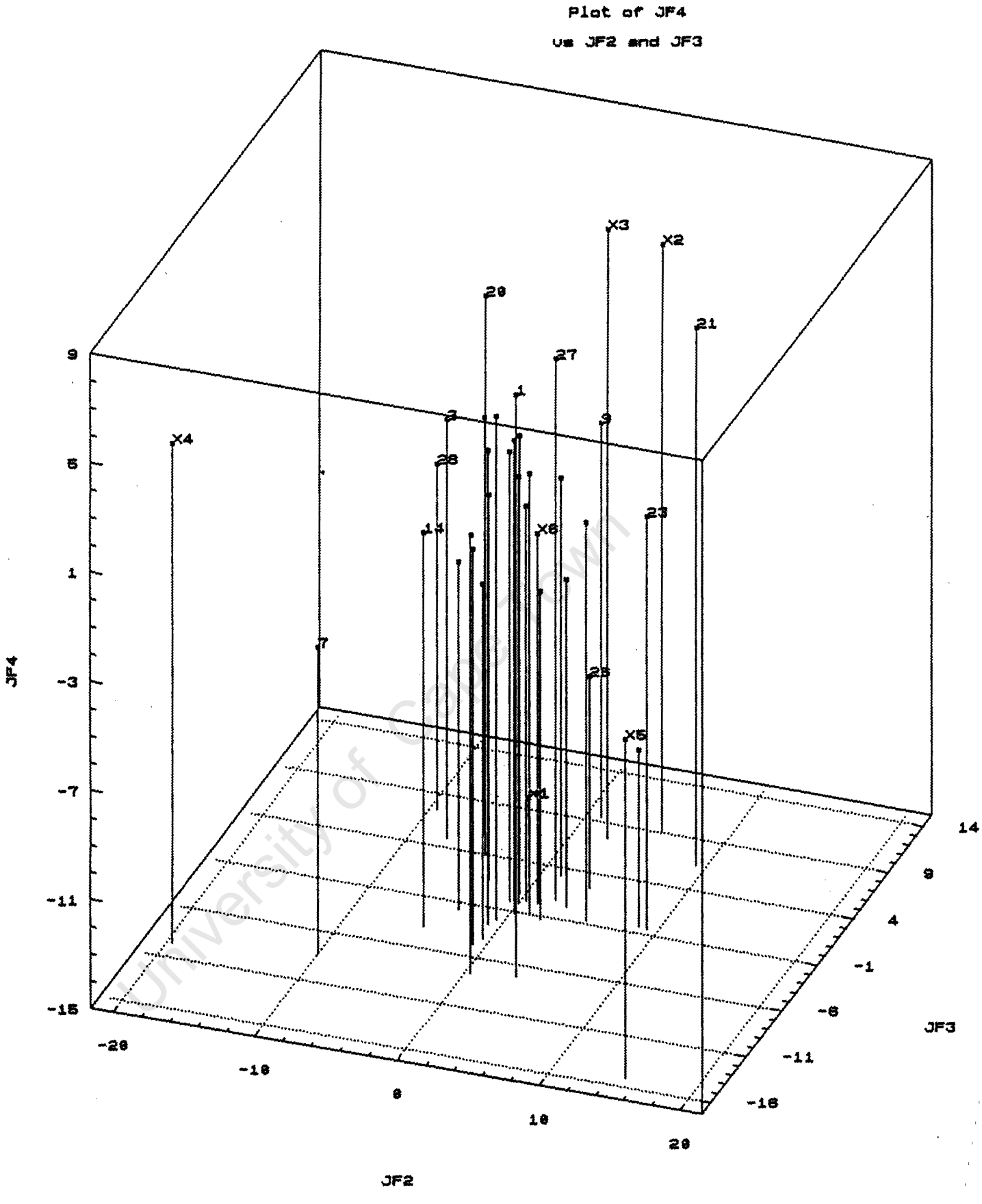
## 4.9 MALLOWS BOUNDED-INFLUENCE ESTIMATORS

The question could be asked whether a more robust approach could not be used to determine on the one hand the influential observations and on the other hand more "stable" estimators for the model.

A bounded-influence type estimator using a weight function was developed by Mallows (1973, 1975). See also De Jogh, De Wet and Welsh (1988). The weights can be constructed in the following way:

Assuming a constant in the model denote the n observations of the (j-1)th independent variable by $x_{1j}, x_{2j},\ldots,x_{nj}$ for $j = 2,3,\ldots,p$. Order the n





observations so that we have $x_{(1)j}, x_{(2)j},\ldots,x_{(n)j}$ and define $r_{1j}, r_{2j},\ldots,r_{nj}$ as the vector of ranks of $x_{1j}, x_{2j},\ldots,x_{nj}$. Let $L = [\tau n] + 1$ and $U = n+1-L$.

The weights associated with the (j-1)th independent variable are now defined as

$$W_{ij} = 1, \quad L \leq r_{ij} \leq U$$

$$= \frac{x_{(L)j} - x_{(U)j}}{D_{ij}}, \quad r_{ij} < L$$

$$= \frac{x_{(U)j} - x_{(L)j}}{D_{ij}}, \quad r_{ij} > U \qquad 4.9.1$$

where

$$D_{ij} = 2x_{ij} - x_{(U)j} - x_{(L)j}, \quad i = 1,\ldots,n$$

The Mallows weights are now defined as

$$W_i = \prod_{j=2}^{p} W_{ij}, \quad i = 1,\ldots,n \qquad 4.9.2$$

Note that the weights are chosen so that the outliers in the independent variable space (influential observations) are given less weight according to their distance from the centre of the independent variable space. In this case the basic idea is to give weight to those $y_i$ having their corresponding $x_i$ in the centre and to decrease weight if $x_i$ is in the tail of the independent variable space. De Jongh et al (1988) chose $\tau = 0.15$ which they showed worked well in

One should see in these results that one should not be too bold in quoting the down weighting factor $\tau$.

When $\tau$ = 0.15 the influential observations are certainly pointed out, but the estimates are affected severely due to the presence of some large y-values which bring more heavy weighting in as a factor at other non-influential points. An additional factor is the fact that the data is simulated. At the end of the next chapter a proposal will be made to cope with this problem in some way.

## 4.10 SUMMARY AND CONCLUSIONS

In this chapter it was shown that the singular value decomposition of the jacobian of a non-linear model can lead to techniques which can give insight into the structure of the data set. This approach leads further to the application of principal components analysis on the so-called "gradiential" data set which in fact, was nothing else than the jacobian. Investigating the principal components which resulted we could pick out and investigate collinearity. Besides this we used the relative variance contributions of rows, i.e. "gradiential" observations to the total variance to indicate possible outliers/influential observations. In a similar way the relative variance contributions of columns can be investigated and small contributions were indicative of "large" correlations or high collinearities involved.

It was also useful to apply the diagonal elements of the hat matrix based on the jacobian as indicators of leverage. This hat matrix was also applied in well known measures like DFFITS[2], replacing the hat matrix based on the data matrix itself.

Taking all these applications together a rather meaningful analysis could be completed which included a graphical display through which one could see how the observations on the outside of the convex hull and the influential observations are the same.

Reference was also made to the Mallows approach to regression with some remarks regarding the careful application of this technique.

Some measures of influence and outliers together with some robust estimation methods were proposed in sections 4.5, 4.6 and 4.7. These proposals are at present under investigation.



## 4.11 EXTENSIONS

Although quite a number of well-known measures were shown to be left open for adaptations like leaving out observations, diminishing rank, and/or using Levenburg-Marquardt changes we did not apply all these options as the interest was mainly aimed at the SVD principal components aspect of the analysis. It can be a full study on its own to find out what the merits are in these applications as indicated in the last part of section 4.5 as well as sections 4.6 and 4.7.

One of merits of these measures is that they relate directly to specific observations and therefore they cannot be ignored, especially when a model structure is such that it can become very difficult to find the relation between variables and columnar results.

A last remark regarding the Mallows estimators is appropriate. If one considers

the weighting process carefully it is obvious that the way it is being done, excludes the non-linear situation as applicable especially with simulated data. It does however point out deviant observations through the weights and in this sense it can be quite useful even in the non-linear situation.

---

distance for uncorrelated linear compounds of the original variates in order to derive their discriminant functions.

For simplicity assume random samples of $n_1$ and $n_2$ observation vectors drawn independently from respective p-dimensional multinormal populations with mean vectors $\mu_1$ and $\mu_2$ and a common var-covariance matrix $\Sigma$. The assumption of normality is not a necessity, but supplies inferential advantages. Construct a linear compound (index) for summarizing observations from the groups on a one-dimensional scale that discriminates between the populations by some measure of maximal separation.

The linear discriminant function then is (see Morrison, 1976)

$$y = (\bar{x}_1 - \bar{x}_2)'S^{-1}x \qquad 5.1.1$$

where $\bar{x}_i$, $i = 1,2$ are the sample mean vectors, $x$ is the observation to be classified and S is the pooled sample var-covariance matrix.

The decision rule is found by determining the average of the discriminant function between the two populations, i.e. where for

$$y_1 = (\bar{x}_1 - \bar{x}_2)'S^{-1}\bar{x}_1 \text{ and } y_2 = (\bar{x}_1 - \bar{x}_2)'S^{-1}\bar{x}_2 \qquad 5.1.2$$

the midpoint

$$c = \tfrac{1}{2}(\bar{x}_1 - \bar{x}_2)S^{-1}(\bar{x}_1 + \bar{x}_2) \qquad 5.1.3$$

serves as classification rule.

Assign observation $\underline{x}$ to population 1 if

$$(\bar{\underline{x}}_1 - \bar{\underline{x}}_2)'S^{-1}\underline{x} > c$$

and to population 2 if

$$(\bar{\underline{x}}_1 - \bar{\underline{x}}_2)'S^{-1}\underline{x} \leq c \ . \qquad 5.1.4$$

This rule can be rewritten as the single statistic of Wald-Anderson, i.e.

$$\begin{aligned} W_{12} &= \underline{x}'S^{-1}(\bar{\underline{x}}_1 - \bar{\underline{x}}_2) - \tfrac{1}{2}(\bar{\underline{x}}_1 - \bar{\underline{x}}_2)'S^{-1}(\bar{\underline{x}}_1 + \bar{\underline{x}}_2) \\ &= -\tfrac{1}{2}(\underline{x} - \bar{\underline{x}}_1)'S^{-1}(\underline{x} - \bar{\underline{x}}_1) + \tfrac{1}{2}(\underline{x} - \bar{\underline{x}}_2)'S^{-1}(\underline{x} - \bar{\underline{x}}_2) \\ &= -\tfrac{1}{2}D_1^2 + \tfrac{1}{2}D_2^2 \ . \end{aligned} \qquad 5.1.5$$

Assign $\underline{x}$ to population 1 if $W_{12} > 0$ and otherwise to population 2, i.e. classify in population 1 when $D_1^2 < D_2^2$, or

$$(\underline{x} - \bar{\underline{x}}_1)'S^{-1}(\underline{x} - \bar{\underline{x}}_1) < (\underline{x} - \bar{\underline{x}}_2)'S^{-1}(\underline{x} - \bar{\underline{x}}_2) \qquad 5.1.6$$

where $D_i^2$, $i = 1,2$ are the Mahalanobis distances for $\underline{x}$ with respect to populations one and two respectively. For a more detailed version and extensions as well as more information on discriminant rules for other data types and distributions, see inter alia Morrison (1976) and Van Deventer (1985).

## 5.2 MEASURING COLLINEARITY

Write the full data matrix and the two population observation matrices as follows:

$$X: n \times p \; ; \quad X_1: n_1 \times p \; ; \quad X_2: n_2 \times p \; ; \quad n_1 + n_2 = n$$

where it is assumed that the population is continuous, measurable on an interval scale and that X is centered and standardized. Note that this does not interfere with the results as the linear discriminant analysis procedure is invariant to standardization.

In practice one often finds that measurements are highly correlated. Obviously, this plays a large role in the results obtained from an analysis, especially in the linear discriminant or quadratic discriminant function (LDF or QDF) techniques of Fisher, for the covariance matrices will tend to be ill-conditioned. Besides this, there is always the probability of the presence of influential observations and/or outliers.



Looking at the LDF as the Mahalanobis distance between the centroids of two population $(\bar{x}_1 - \bar{x}_2)'S^{-1}(\bar{x}_1 - \bar{x}_2)$ it is obvious that the inverted var-cov. matrix must be of such a nature that it is not ill-conditioned in any way and certainly not singular. Even if the observations are measured accurately, the minimum mean error squared of coefficients in the discriminant function can not be guaranteed, because of rounding off errors, intercorrelation of the independent variables which may give rise to ill-conditioned situations as well as a too large a number of variables which may lead to "dummy" intercorrelation.

As sample sizes are often small, biasing could probably be used whenever classification problems arise. Biasing most greatly affects the variables corresponding to the smallest eigenvalues of the var-covariance matrix $\Sigma$. However, $\Sigma$ is usually unknown and the effectiveness of biasing will have to be determined from the samples.

Before one starts off on biasing or other techniques to correct for intercorrelation etc, one must first try to determine whether there really is a problem. Several authors (see Forsythe and Moler, 1967; Marshall and Olkin, 1968; Vinod, 1976) proposed the use of the "condition number" to measure the instability of a matrix when solving for a system a linear equations (Troskie, 1977). The distribution of the ratios of characteristic roots (condition numbers) and their applications are also of importance (Troskie and Conradie, 1986). The condition number is usually defined as

$$C_Q = Q(A).Q(A^{-1}) \qquad 5.2.1$$

where Q is usually taken as the norm. In the linear model the condition number for A (i.e. X'X) is

$$C_Q = C(X'X) = \frac{\lambda_1}{\lambda_p} \qquad 5.2.2$$

where $\lambda_1 > \ldots > \lambda_p$, the characteristic roots of X'X.

The condition number is a better measure of the nearness to singularity than the determinant of A. According to Belsley, Kuh and Welsch (1980) a condition number

of more than 10 indicates weak dependencies. A condition number of 15-30 has an associated correlation coefficient of 0,9 and a condition number of 100 and more shows serious problems in the solutions. A rule of thumb may be that condition numbers shouldn't be much larger than 10p, where p is the number of unknown parameters. The condition index as defined by Belsley, Kuh and Welsch (1980) is defined as $\sqrt{\lambda_1/\lambda_p}$. Better defined definitions of what is a "large" condition number can be found in Troskie and Conradie (1986). Troskie and Conradie considered also measures of condition like $\Sigma\lambda_i/\lambda_p$ and more specifically $tr(X'X)/p$.

## 5.3 SINGULAR VALUE DECOMPOSITION AND BIASING IN THE PRESENCE OF COLLINEARITY

The application of bias to discriminant analysis was discussed inter alia by DiPillo (1979). His discussion was based on a reduction in variance and the effect of that on the probability of misclassification. DiPillo did investigate one of the procedures described in this paragraph, i.e. the use of S + kI in stead of S. He mentioned the use of S + kD, but no results have appeared as far as the author is aware of. In this section several methods are employed to indicate in which way a reduction in variance may be obtained through a bias in the singular value decomposition, in the first place on the pooled sample variance S or the correlation matrix R and in the second place in a more direct way on the data matrix X as well as $X_1$ and $X_2$ as defined in 5.2. For more information on the basic structure display of a data matrix and discriminant analysis refer to appendix B.

### 5.3.1 Bias on the var-covariance and/or correlation matrix through SVD

Let $\underline{x}' = (x_1,\ldots,x_p)$ be an observation from one of two p-variate normal populations, say $\Pi_i \sim n(\mu_i, p, \Sigma_i)$, i = 1,2. Assume that $\Sigma_1 = \Sigma_2 = \Sigma$. The pooled sample var-cov. matrix is S. Write S in its singular value decomposition form, i.e. (using V instead of D for the diagonal matrix for notational purposes - see 5.3.7)

$$S = U_s V_s V_s' \qquad 5.3.1.1$$

The interest now lay in applying a bias on the diagonal in V, i.e.

$$S^* = U_s(V_s + k_s I)V_s', \quad k_s > 0 \qquad 5.3.1.2$$

$$\text{If } Q_j(\underline{x}) = (\underline{x} - \bar{\underline{x}}_j)'S^{-1}(\underline{x} - \bar{\underline{x}}_j), \; j = 1,2 \qquad 5.3.1.3$$

with rule: If $\min_j[Q_j(\underline{x})] = Q_i(\underline{x})$, classify $\underline{x}$ into $\Pi_i$, one can write from 3.1.1, 3.1.2 and 3.1.3

$$\begin{aligned} {}_sQ_j^*(\underline{x}) &= (\underline{x} - \bar{\underline{x}}_j)'(U_s(V_s + k_s I)V_s')^{-1}(\underline{x} - \bar{\underline{x}}_j) \\ &= (\underline{x} - \underline{x}_j)'(U_s V_s V_s' + k_s U_s V_s')^{-1}(\underline{x} - \bar{\underline{x}}_j) \\ &= (\underline{x} - \bar{\underline{x}}_j)'(S + k_s U_s V_s')^{-1}(\underline{x} - \bar{\underline{x}}_j) \qquad 5.3.1.4 \end{aligned}$$

Another approach could be to write

$$S = D^{\frac{1}{2}}RD^{\frac{1}{2}} \qquad 5.3.1.5$$

with R the correlation matrix, and further

$$R = V_R W_R V_R' \qquad 5.3.1.6$$

i.e.

$$S = D^{\frac{1}{2}}(V_R W_R V_R')D^{\frac{1}{2}} \qquad 5.3.1.7$$

and

$$S_R^* = D^{\frac{1}{2}}(V_R(W_R + k_R I)V_R')D^{\frac{1}{2}}, \quad k_R > 0 \qquad 5.3.1.8$$

so that

$$\begin{aligned} {}_R Q_j^*(\underline{x}) &= (\underline{x} - \bar{\underline{x}}_j)'(D^{\frac{1}{2}}(V_R(W_R + k_R I)V_R')D^{\frac{1}{2}})^{-1}(\underline{x} - \bar{\underline{x}}_j) \\ &= (\underline{x} - \bar{\underline{x}}_j)'(S + D^{\frac{1}{2}}V_R k_R I V_R' D^{\frac{1}{2}})^{-1}(\underline{x} - \bar{\underline{x}}_j) \\ &= (\underline{x} - \bar{\underline{x}}_j)'(S + k_R D)^{-1}(\underline{x} - \bar{\underline{x}}_j) \ . \end{aligned} \qquad 5.3.1.9$$

Let ${}_S G^*(\underline{x}) = {}_S Q_1^*(\underline{x}) - {}_S Q_2^*(\underline{x})$

and ${}_R G^*(\underline{x}) = {}_R Q_1^*(\underline{x}) - {}_R Q_2^*(\underline{x})$.

If ${}_S G^*(\underline{x})$ or ${}_R G^*(\underline{x}) < 0$ classify in population 1, otherwise in population 2.

Note that ${}_{S}G^{*}(\underline{x})$ and ${}_{R}G^{*}(\underline{x})$ can be simplified to

$$ {}_{S}G^{*}(\underline{x}) = (\bar{\underline{x}}_2 - \bar{\underline{x}}_1)'(S + k_s U_s V_s')^{-1}\underline{x} - \tfrac{1}{2}(\bar{\underline{x}}_2 - \bar{\underline{x}}_1)'(S + k_s U_s V_s')^{-1}(\bar{\underline{x}}_2 + \bar{\underline{x}}_1) \qquad 5.3.1.10 $$

and

$$ {}_{R}G^{*}(\underline{x}) = (\bar{\underline{x}}_2 - \bar{\underline{x}}_1)'(S + k_R D)^{-1}\underline{x} - \tfrac{1}{2}(\bar{\underline{x}}_2 - \bar{\underline{x}}_1)'(S + k_R D)^{-1}(\bar{\underline{x}}_2 + \bar{\underline{x}}_1) \qquad 5.3.1.11 $$

Given that $\underline{x}$ is from $\Pi_1$ it can be shown that

$$ \begin{aligned} G(\underline{x}) &= Q_1(\underline{x}) - Q_2(\underline{x}) \\ &= (\underline{x} - \bar{\underline{x}}_1)'S^{-1}(\underline{x} - \bar{\underline{x}}_1) - (\underline{x} - \bar{\underline{x}}_2)'S^{-1}(\underline{x} - \bar{\underline{x}}_2) \end{aligned} \qquad 5.3.1.12 $$

is distributed normal with

$$ E\left[G(\underline{x})/\bar{\underline{x}}_1,\bar{\underline{x}}_2,S,\Pi_1\right] = (\bar{\underline{x}}_2 - \bar{\underline{x}}_1)'S^{-1}\mu_1 - \tfrac{1}{2}(\bar{\underline{x}}_2 - \bar{\underline{x}}_1)S^{-1}(\bar{\underline{x}}_2 + \bar{\underline{x}}_1) \qquad 5.3.1.13 $$

and

$$ \mathrm{var}\left[G(x)/\bar{\underline{x}}_1,\bar{\underline{x}}_2,S,\Pi_1\right] = (\bar{\underline{x}}_2 - \bar{\underline{x}}_1)'S^{-1}\Sigma S^{-1}(\bar{\underline{x}}_2 - \bar{\underline{x}}_1) \qquad 5.3.1.14 $$

Similar results apply for ${}_{S}G^{*}(\underline{x})$ and ${}_{R}G^{*}(\underline{x})$. It can be shown that ${}_{S}G^{*}(\underline{x})$ and ${}_{R}G^{*}(\underline{x})$ are normally distributed with

$$ E\left[{}_{S}G^{*}(\underline{x})/\bar{\underline{x}}_1,\bar{\underline{x}}_2,S,\Pi_1\right] = (\bar{\underline{x}}_2 - \bar{\underline{x}}_1)'(S + k_s U_s V_s')^{-1}\mu_1 $$

$$- \tfrac{1}{2}(\bar{x}_2 - \bar{x}_1)(S + k_s U_s V_s')^{-1}(\bar{x}_2 + \bar{x}_1) \qquad 5.3.1.15$$

and

$$\text{var}\left[{}_S G^*(x)/\bar{x}_1,\bar{x}_2,S,\Pi_1\right] = (\bar{x}_2 - \bar{x}_1)'(S + k_s U_s V_s')^{-1}\Sigma(S + k_s U_s V_s')^{-1}(\bar{x}_2 - \bar{x}_1) \qquad 5.3.1.16$$

with corresponding results for ${}_R G^*(x)$ i.e.

$$E\left[{}_R G^*(x)/\bar{x}_1,\bar{x}_2,S,\Pi_1\right] = (\bar{x}_2 - \bar{x}_1)'(S + k_R D)^{-1}\mu_1 - \tfrac{1}{2}(\bar{x}_2 - \bar{x}_1)(S + k_R D)^{-1}(\bar{x}_2 + \bar{x}_1) \qquad 5.3.1.17$$

and

$$\text{var}\left[{}_R G^*(x)/\bar{x}_1,\bar{x}_2,S,\Pi_1\right] = (\bar{x}_2 - \bar{x}_1)'(S + k_R D)^{-1}\Sigma(S + k_R D)^{-1}(S + k_R D)^{-1}(\bar{x}_2 - \bar{x}_1) \qquad 5.3.1.18$$

It can further be shown that

$$\text{Var}\left[G(x)/\bar{x}_1,\bar{x}_2,S,\Pi_1\right] > \text{Var}\left[{}_{R \text{ or } S} G^*(x)/\bar{x}_1,\bar{x}_2,S,\Pi_1\right], \quad k_{s \text{ or } R} > 0 \qquad 5.3.1.19$$

The shift in location from $G(x)$ to ${}_R G^*(x)$ or ${}_S G^*(x)$ may cause inconsistent results. This biasedness, however may be overcome consistently by the reduction in variance of the classification rule.

The probability of misclassification in the case where all population parameters

are known is $\phi(-\frac{\delta}{2})$ with

$$\delta^2 = (\mu_2 - \mu_1)'\Sigma^{-1}(\mu_2 - \mu_1) \qquad 5.3.1.20$$

In the case where only sample estimates are available the total probability of misclassification (PMC) is

$$\tfrac{1}{2}(1 - \phi(z_1) + \phi(z_2)) \qquad 5.3.1.21$$

where

$$z_i = \frac{\frac{1}{2}(\bar{x}_2 - \bar{x}_1)'S^{-1}(\bar{x}_2 + \bar{x}_1) - (\bar{x}_2 - \bar{x}_1)'\mu_i}{[(\bar{x}_2 - \bar{x}_1)'S^{-1}\Sigma S^{-1}(\bar{x}_2 - \bar{x}_1)]^{\frac{1}{2}}}; \quad i = 1,2 \qquad 5.3.1.22$$

Under the biased SVD correlation coefficient matrix or biased SVD variance covariance matrix, let the biased probabilities of misclassification (PMCR* and PMCS*) be

$$\tfrac{1}{2}(1 - \phi({}_R z_1^*) + \phi({}_R z_2^*)) \text{ and } \tfrac{1}{2}(1 - \phi({}_S z_1^*) + \phi({}_S z_2^*)) \qquad 5.3.1.23$$

where for $k > 0$

$$S^{z_i^*} = \frac{\frac{1}{2}(\bar{x}_2 - \bar{x}_1)'(S + k_s U_s V_s')^{-1}(\bar{x}_2 + \bar{x}_1) - (\bar{x}_2 - \bar{x}_1)'(S + k_s U_s V_s')^{-1}\mu_i}{\left[(\bar{x}_2 - \bar{x}_1)'(S + k_s U_s V_s')^{-1}\,\Sigma(S + k_s U_s V_s')^{-1}(\bar{x}_2 - \bar{x}_1)\right]^{\frac{1}{2}}}; \quad i=1,2 \qquad 5.3.1.24$$

and

$$ {}_R^{*}z_i = \frac{\frac{1}{2}(\bar{x}_2-\bar{x}_1)'(S+k_R D)^{-1}(\bar{x}_2+\bar{x}_1) - (\bar{x}_2-\bar{x}_1)'(S+k_R D)^{-1}\mu_i}{\left[(\bar{x}_2-\bar{x}_1)'(S+k_R D)^{-1}\,\Sigma(S+k_R D)^{-1}\,(\bar{x}_2-\bar{x}_1)\right]^{\frac{1}{2}}}; \quad i=1,2 $$

5.3.1.25

Use the sign of the rate of change of PMCR* and PMCS* to determine whether the probability of misclassification according to either rule increases or decreases for a given k > 0. So from 3.1.18

$$ \frac{d}{dk}(PMCS^{*}) = -\frac{e^{-\frac{1}{2}({}_S^{*}z_1)^2}}{\sqrt{2\pi}}\frac{d_S{}^{*}z_1}{dk} + \frac{e^{-\frac{1}{2}({}_S^{*}z_2)^2}}{\sqrt{2\pi}}\frac{d_S{}^{*}z_2}{dk} $$

5.3.1.26

with

$$ \frac{d_S{}^{*}z_i}{dk} = \frac{\frac{1}{2}(\bar{x}_2-\bar{x}_1)'(S+k_s U_s V_s')^{-1}(\bar{x}_2+\bar{x}_1-2\mu_i)\cdot(\bar{x}_2-\bar{x}_1)'(S+k_s U_s V_s')^{-2}\Sigma(S+k_s U_s V_s')^{-1}(\bar{x}_2-x_1)}{\left[(\bar{x}_2-\bar{x}_1)'(S+k_s U_s V_s')^{-1}\,\Sigma(S+k_s U_s V_s')^{-1}\,(\bar{x}_2-\bar{x}_1)\right]^{\frac{3}{2}}} $$

$$ - \frac{\frac{1}{2}(\bar{x}_2-\bar{x}_1)'(S+k_s U_s V_s')^{-2}(\bar{x}_2+\bar{x}_1-2\mu_i)}{\left[(\bar{x}_2-\bar{x}_1)'(S+k_s U_s V_s')^{-1}\,\Sigma(S+k_s U_s V_s')^{-1}\,(\bar{x}_2-\bar{x}_1\right]^{\frac{1}{2}}}, \quad i=1,2 $$

5.3.1.27

and

$$ \frac{d}{dk}(PMCR^{*}) = -\frac{e^{-\frac{1}{2}({}_R^{*}z_1)^2}}{\sqrt{2\pi}}\frac{d_R{}^{*}z_1}{dk} + \frac{e^{-\frac{1}{2}({}_R^{*}z_2)^2}}{\sqrt{2\pi}}\frac{d_R{}^{*}z_2}{dk} $$

5.3.1.28

with

$$\frac{d_R z_i^*}{dk} = \frac{\frac{1}{2}(\bar{x}_2-\bar{x}_1)'(S+k_R D)^{-1}(\bar{x}_2+\bar{x}_1-2\mu_i)\cdot(\bar{x}_2-\bar{x}_1)'(S+k_R D)^{-2}\Sigma(S+k_R D)^{-1}(\bar{x}_2-\bar{x}_1)}{\left[(\bar{x}_2-\bar{x}_1)'(S+k_R D)^{-1}\ \Sigma(S+k_R D)^{-1}(\bar{x}_2-\bar{x}_1)\right]^{\frac{3}{2}}}$$

$$- \frac{\frac{1}{2}(\bar{x}_2-\bar{x}_1)'(S+k_R D)^{-2}(\bar{x}_2+\bar{x}_1-2\mu_i)}{\left[(\bar{x}_2-\bar{x}_1)'(S+k_R D)^{-1}\ \Sigma(S+k_R D)^{-1}\ (\bar{x}_2-\bar{x}_1)\right]^{\frac{1}{2}}}\ ,\ i=1,2 \qquad 5.3.1.29$$

When $k = 0^+$, more specifically $0 < k < 1$, negative derivatives in 5.3.1.21 and 5.3.1.23 imply that the introduction of k as a bias leads to a reduction in the probability of misclassification.

DiPillo (1979) found a marked increase in efficiency using computer runs on his biased procedure S+kI, not so much when the sample sizes are large, but especially when $n$ is small and or the number of variables increases.

It should be illuminating applying these new biased procedures on DiPillo's data sets.

The optimum value of $k_s$ or $k_R$ is difficult to determine. It should obviously be where $\frac{d(PMCS^*)}{dk}$ or $\frac{d(PMCR^*)}{dk}$ equals zero - not an easy equation to solve. Other methods depend on the maximisation of

$$S\Delta_*^2 = (\bar{x}_1 - \bar{x}_2)'(S+k_s U_s V_s')^{-1}(\bar{x}_1 - \bar{x}_2) \qquad 5.3.1.30$$

or

$$R\Delta_*^2 = (\bar{x}_1 - \bar{x}_2)'(S+k_R D)^{-1}(\bar{x}_1 - \bar{x}_2) \qquad 5.3.1.31$$

as well as examination of reversal in direction of individual coefficients in the discriminant function.

### 5.3.2 Bias on the data matrix and SVD

Assume that the matrix X:n×p; $X_1$: $n_1$×p and $X_2$: $n_2$×p, $n_1+n_2 = n$ are standardized. Let

$$X_1 = U_1 D_{1a} V_1' \text{ and } X_2 = U_2 D_{2a} V_2' \qquad 5.3.2.1$$

In section 5.3.1 a bias was introduced on the var-covariance matrix or the correlation matrix through the SVD when collinearity was present. I propose a bias on the group data matrices themselves in the Levenburg-Marquardt way in the presence of collinearity:

$$X_{1L} = U_1(D_{1a} + \lambda_1 I)V_1' \text{ and } X_{2L} = U_2(D_{2a} + \lambda_2 I)V_2' \qquad 5.3.2.2$$

Combine these adapted group data matrices in $X_L$, the Levenburg-Marquardt adapted total data matrix and carry out the discriminant analysis in the usual way using e.g. the LDF of Fisher, comparing results for different pairs of values of $\lambda_1$ and $\lambda_2$.

Another way of approaching the problem is to use the r columns and rows associated with the r largest roots in each of the group data matrices, i.e. let

$$X_{1r} = U_{1[r]} D_{1a[r]} V_{1[r]}' \text{ and } X_{2r} = U_{2[r]} D_{2a[r]} V_{2[r]}' \qquad 5.3.2.3$$

Combine these adapted group data matrices in $X_{[r]}$, the diminished rank total data matrix and carry out the discriminant analysis in the usual way using again e.g. the LDF of Fisher.

These two approaches can be put together by creating

$$X_{1Lr} = U_{1[r]}(D_{1a[r]} + \lambda_{1r})V'_{1[r]}, \text{ and}$$

$$X_{2Lr} = U_{2[r]}(D_{2a[r]} + \lambda_{2r})V'_{2[r]} \qquad 5.3.2.4$$

Combine these adapted group data matrices in $X_{Lr}$ the Levenburg-Marquardt diminished rank total data matrix and carry out the discriminant analysis in the usual way using again e.g. the LDF of Fisher.

Using these different approaches a reduction in misclassification error rate may be obtained as well as a more efficient and less complicated discrimination function. It is also obvious that these techniques can easily be extended to more than two populations.



## 5.4 OUTLIER DETECTION IN DISCRIMINANT ANALYSIS USING THE INFLUENCE FUNCTION

The influence function can be used as an aid in outlier detection in discriminant analysis. One could determine a statistic based on the group as a whole and also with suspect observation(s) removed or down-scaled. The influence of this procedure on the statistic can then be determined.

In discriminant analysis there is a variety of statistics which one can use in

the influence function, inter alia Mahalanobis's $\Delta^2$, a function of the coefficient vector of the discriminant function and the group means.

Campbell (1978) gave a number of references in this respect and then carried on to apply the influence function as an aid to outlier detection in discriminant analysis.

He distinguished between the theoretical and sample influence functions. Empirically it focuses on $\theta - \theta_{-m}$ where $\theta$ is an estimator based on n observations and $\theta_{-m}$ is an estimator, similar to $\theta$, but determined without the mth observation. The influence function is then given by

$$I_m(\underline{x},\theta) = \lim_{\varepsilon \to 0} \frac{\hat{\theta}_{-m} - \theta}{\varepsilon} \qquad 5.4.1$$

where $\varepsilon$ is taken as $-\frac{1}{n-1}$, i.e.

$$I_m(\underline{x}, \theta) = (n-1)(\hat{\theta} - \hat{\theta}_{-m}) \qquad 5.4.2$$

Theoretically the perturbed distribution $\bar{F}_k$, where the parameter $\theta = T(F_1,\ldots,F_k,\ldots,F_g)$, may be expressed as

$$\bar{F}_k = (1 - \varepsilon)F_k + \varepsilon\delta_{\underline{x}} \qquad 5.4.3$$

$\delta_{\underline{x}}$ being the distribution function which assigns unit probability to the point $\underline{x}$.

Note that as k refers to the group, we can ignore k and work with one group, say the first only, i.e. we will concentrate on $\bar{F}_1 = (1-\varepsilon)F_1 + \varepsilon\delta_{\underline{x}}$.

Assume a two population discriminant function $(\underline{\mu}_1 - \underline{\mu}_2)'\Sigma^{-1}\underline{x} = \underline{\ell}'\underline{x}$ where $\underline{x} \sim n(p, \underline{\mu}_i, \Sigma)$. Given $\underline{\delta} = \underline{\mu}_1 - \underline{\mu}_2$, we have Mahalanobis's theoretical $\Delta^2 = \underline{\delta}'\Sigma^{-1}\underline{\delta}$ and the discriminant function coefficient vector $\underline{\ell} = \underline{\delta}'\Sigma^{-1}$.

In order to determine the influence functions of $\Delta^2$ and $\underline{\ell}$, we must determine the influence of perturbing for $\underline{x}_m = \underline{x}$ on $\underline{\mu}_i$, $\underline{\delta}$, $\Sigma$ and $\Sigma^{-1}$. Let

$$\Sigma = w_1\Sigma_{F_1} + w_2\Sigma_{F_2}, \qquad w_1 + w_2 = 1, \qquad w_k > 0 \tag{5.4.4}$$

with

$$\Sigma_{F_1} = \int(\underline{x} - \underline{\mu}_i)(\underline{x} - \underline{\mu}_i)'dF_i, \qquad \underline{\mu}_i = \int\underline{x}df_i \tag{5.4.5}$$

For further derivations assume $\Sigma_{F_1} = \Sigma_{F_2}$, so that the weighting factors make provision for unequal sample sizes only. So, if $\rightarrow$ indicates the perturbed parameter

$$\begin{aligned}\underline{\mu}_1 \rightarrow (1 - \varepsilon)\underline{\mu}_1 + \varepsilon\underline{x} &= \underline{\mu}_1 + \varepsilon(\underline{x} - \underline{\mu}_1) \\ &= \underline{\mu}_1 + \varepsilon\underline{z}\end{aligned} \tag{5.4.6}$$

where $\underline{z} = \underline{x} - \underline{\mu}_1$ .

$$\underline{\delta} \rightarrow (1 - \varepsilon)\underline{\delta} + \varepsilon(\underline{x} - \mu_2), \qquad 5.4.7$$

i.e. in the difference between the centroids, $\underline{x}$ will appear with probability 1, in stead of $\mu_1$. So

$$\underline{\delta} \rightarrow (1 - \varepsilon)(\mu_1 - \mu_2) + \varepsilon(\underline{x} - \mu_2) = \mu_1 - \mu_2 + \varepsilon(\underline{x} - \mu_1)$$
$$= \underline{\delta} + \varepsilon\underline{z} \qquad 5.4.8$$

$$\Sigma_{F_1} \rightarrow (1 - \varepsilon)\Sigma_{F_1} + \varepsilon\underline{z}\underline{z}' \qquad 5.4.9$$

With the same reasoning as before 5.4.8 we have $\underline{z}\underline{z}'$ in stead of $\Sigma_{F_1}$, because $\underline{x} - \mu_1 = \underline{z}$ appears with probability 1. Further, keeping in mind that $\Sigma_{F_1} = \Sigma_{F_2}$ and that a weighting factor $w_1$ is involved where $\Sigma = w_1\Sigma_{F_1} + w_2\Sigma_{F_2}$

$$\Sigma \rightarrow (1 - \varepsilon w_1)\Sigma + \varepsilon w_1\underline{z}\underline{z}' \qquad 5.4.10$$

From Press (1972) follows now that

$$\Sigma^{-1} \rightarrow (1 - \varepsilon w_1)^{-1}\left(\Sigma^{-1} - \frac{\varepsilon w_1\Sigma^{-1}\underline{z}\underline{z}'\Sigma^{-1}}{1 - \varepsilon w_1 + w_1\underline{z}'\Sigma^{-1}\underline{z}}\right)$$

$$= (1 + \varepsilon w_1)\Sigma^{-1} - \varepsilon w_1\Sigma^{-1}\underline{z}\underline{z}'\Sigma^{-1} \qquad 5.4.11$$

Now we can determine the influence functions with respect to Mahalanobis's theoretical $\Delta^2$ as well as the discriminant function coefficient vector $\underline{\ell}$, i.e. $I(\underline{x}, \Delta^2)$ and $I(\underline{x}, \underline{\ell})$.

(i) $I(\underline{x}, \Delta^2)$:

From 5.4.8 and 5.4.11 we have

$$\begin{aligned}\Delta^2 &= (\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2)\\ &= \underline{\delta}'\Sigma^{-1}\underline{\delta}\\ &\to (\underline{\delta} + \varepsilon\underline{z})'\{(1 + \varepsilon w_1)\Sigma^{-1} - \varepsilon w_1\Sigma^{-1}\underline{z}\underline{z}'\Sigma^{-1}\}(\underline{\delta} + \varepsilon\underline{z}) \qquad 5.4.12\end{aligned}$$

Let

$$\phi = (\mu_1 - \mu_2)\Sigma^{-1}(\underline{x} - \mu_1) = \underline{\delta}'\Sigma^{-1}\underline{z} \qquad 5.4.13$$

Then, ignoring terms in $\varepsilon$ where the order is more than one it follows from 5.4.12 and 5.4.13 that

$$\begin{aligned}\Delta^2 &\to \underline{\delta}'\Sigma^{-1}\underline{\delta} + \varepsilon w_1\underline{\delta}'\Sigma^{-1}\underline{\delta} + \varepsilon\underline{\delta}'\Sigma^{-1}\underline{z} + \varepsilon\underline{z}'\Sigma^{-1}\underline{\delta} - \varepsilon w_1\underline{\delta}\Sigma^{-1}\underline{z}\underline{z}'\Sigma^{-1}\underline{\delta}\\ &= \Delta^2 + \varepsilon w_1\Delta^2 + \varepsilon\phi + \varepsilon\phi - \varepsilon w_1\phi\phi'\\ &= \Delta^2(1 + \varepsilon w_1) + 2\varepsilon\phi - \varepsilon w_1\phi^2 \qquad 5.4.14\end{aligned}$$

From this the influence function for $\Delta^2$ can be determined, viz. (see 5.4.3 for $\bar{F}$)

$$\begin{aligned}I(\underline{x}, \Delta^2) &= \lim_{\varepsilon\to 0}\frac{\overline{\Delta^2} - \Delta^2}{\varepsilon}\\ &= \lim_{\varepsilon\to 0}\frac{\Delta^2\varepsilon w_1 + 2\varepsilon\phi - \varepsilon w_1\phi^2}{\varepsilon}\\ &= \Delta^2 w_1 + 2\phi - w_1\phi^2 \qquad 5.4.15\end{aligned}$$

The coefficient vector can be standardised using $\sqrt{\Delta^2}$, so that the standardised vector $\underline{\ell}_s$ can be written as $\underline{\ell}_s = \Delta^{-1}\underline{\ell}$. Similarly $\phi$ can be standardised so that the standardised form of $\phi$ is $\phi_s = \Delta^{-1}\phi$. Now $\phi$ is distributed $N(0, \Delta^2)$, i.e. $\phi_s \sim N(0,1)$. Then 5.4.15 in terms of $\phi_s$ becomes:

$$I(\underline{x}, \Delta^2) = \Delta^2 w_1 + 2\Delta\phi_s - w_1\Delta^2\phi_s^2 \qquad 5.4.16$$

(ii) $I(\underline{x}; \underline{\ell})$:

In the same way as before and using 5.4.8 and 5.4.11

$$\begin{aligned}\underline{\ell} = \Sigma^{-1}\underline{\delta} &\rightarrow ((1 + \epsilon w_1)\Sigma^{-1} - \epsilon w_1\Sigma^{-1}\underline{z}\underline{z}'\Sigma^{-1})(\underline{\delta} + \epsilon\underline{z}) \\ &= (1 + \epsilon w_1)\Sigma^{-1}\underline{\delta} - \epsilon w_1\Sigma^{-1}\underline{z}\underline{z}'\Sigma^{-1}\underline{\delta} + (1 + \epsilon w_1)\Sigma^{-1}\epsilon\underline{z} - \epsilon w_1\Sigma^{-1}\underline{z}\underline{z}'\Sigma^{-1}\epsilon\underline{z} \\ &= \underline{\ell} + \epsilon w_1\underline{\ell} - \epsilon w_1\Sigma^{-1}\underline{z}\phi + \Sigma^{-1}\epsilon\underline{z} \qquad 5.4.17\end{aligned}$$

ignoring terms with $\epsilon^2$ and smaller factors and taking 5.4.13 into account. Then the influence function is

$$\begin{aligned}I(\underline{x}; \underline{\ell}) &= \lim_{\epsilon\to 0}\frac{\underline{\ell} - \underline{\ell}}{\epsilon} = \lim_{\epsilon\to 0}\frac{(\underline{\ell} + \epsilon w_1\underline{\ell} - \epsilon w_1\Sigma^{-1}\underline{z}\phi + \Sigma^{-1}\epsilon z) - \underline{\ell}}{\epsilon} \\ &= w_1\underline{\ell} - w_1\Sigma^{-1}\underline{z}\phi + \Sigma^{-1}\underline{z} \\ &= w_1\underline{\ell} + (1 - w_1\phi)\Sigma^{-1}\underline{z} \qquad 5.4.18\end{aligned}$$

And as before we can determine $I(\underline{x}; \underline{\ell}_s)$ and also $I(\underline{x}; \underline{\ell}'\underline{\ell})$; where $\underline{\ell}_s$ refers to the scaled vector:

$$I(\underline{x}; \underline{\ell}_s) = \{\tfrac{1}{2}w_1 - \tfrac{1}{2}w_1\phi(2 - \phi)\Delta^{-2}\}\underline{\ell}_s - \Delta^{-1}\Sigma^{-1}\underline{z}(w_1\phi - 1) \qquad 5.4.19$$

$$I(\underline{x};\ \underline{\ell}'\underline{\ell}) = 2w_1\underline{\ell}'\underline{\ell} + 2(1 - w_1\phi)\underline{\ell}'\Sigma^{-1}\underline{z} \qquad 5.4.20$$

In the sample analogues of the theoretical influence functions $\epsilon$ is replaced by $\frac{-1}{n-1}$ and terms of order $(n^{-2})$ are ignored. Further $\phi_s$ is replaced by $\text{est}(\phi_s) = \text{est}(\Delta^{-1}\phi) = \text{est}(\Delta^{-1}\underline{\ell}(\underline{x}_m - \mu_1)) = \underline{c}_s'(\underline{x}_m - \bar{\underline{x}}_1)$ where $\underline{c}_s$ is the standardised vector of the sample discriminant function coefficients, $w_k = \frac{n_k}{n_1+n_2}$, $\underline{x}_m$ is the mth observation and $\text{est}(\cdot) \equiv$ estimator of $(\cdot)$. Mahalanobis's $\Delta^2$ is replaced by $D^2$ and $\underline{\ell}'\Sigma^{-1}(\underline{x}_m - \mu_1)$ is replaced by $(\bar{\underline{x}}_1 - \bar{\underline{x}}_2)'S^{-1}(\underline{x}_m - \bar{\underline{x}}_1)$; S being the unbiased pooled cov. matrix based on $n_1 + n_2 - 2$ degrees of freedom.

In an application on real data Campbell plotted the change in $D^2$ against the standardised discriminant score as a deviation from that for the species mean, i.e. he plotted $D^2 - D^2_{-m}$ against $\underline{c}_s'(\underline{x}_m - \bar{\underline{x}}_1)$. If one uses now the asymptotically normal distribution of the discriminant scores one can decide on possible outliers.

A second technique is to plot $D^2 - D^2_{-m}$ against a gamma distribution with parameters estimated by means of the maximum likelihood method from the smallest 95 order statistics. In fact Campbell didn't use $D^2 - D^2_{-m}$ as such, but

$$\begin{aligned} I_M(\underline{x},\ D^2) &= I_{Max}(\underline{x},\ D^2) - I(\underline{x},\ D^2) \\ &= w_1^{-1}(1 - 2w_1 D\underline{c}_s'(\underline{x}_m - \bar{\underline{x}}_1) + w_1^2D^2\left[\underline{c}_s'(\underline{x}_m - \bar{\underline{x}}_1)\right]^2) \\ &\doteq w_1D^2(\underline{c}_s'(\underline{x}_m - \bar{\underline{x}}_1) - w_1^{-1}D^{-1})^2 \end{aligned} \qquad 5.4.21$$

$\doteq \chi^{2'}$ with 1 degree of freedom and non-centrality parameter $(w_1^2D^2)^{-1}$ which suggests a gamma distribution. Campbell compared the moments for $I_m(\underline{x},\ D^2)$ with

those for the $c\chi^2_\nu$ distribution, where $c = \nu^{-1}(1 + w_1^{-2}D^{-2})$ and $\nu = 1 + (2w_1^2D^2 + w_1^4D^4)^{-1}$.

In a similar way he used a third plot, viz. that of

$$I_M(\underline{x};\ \underline{c}'\underline{c}) = 2(1 - w_1 D\underline{c}'_S(\underline{x}_m - \bar{\underline{x}}_1))\underline{c}'S^{-1}(\underline{x}_m - \bar{\underline{x}}_1) \qquad 5.4.22$$

against the gamma quantiles as in the previous case.

All these techniques lead to the same indication of outliers in his example.

The study has shown that observations do influence $D^2$ rather assymetrically and also that the inclusion of an observation lying further from the mean of the other group decreases rather than increases $D^2$. All in all Campbell came to the conclusion that probability plots of the $D^2 - D^2_m$ values against the appropriate gamma distribution would seem to be preferable to probability plots of the discriminant scores against expected normal order statistics or similar appropriate normal plotting positions.

The drawback in the procedures mentioned above is the immense amount of processing necessary to make the system work. I would like to propose a different way of looking at the data matrix - as a whole, but also the group data matrices. Using the singular value decomposition and the variance ratios as in the previous chapters to decide on possible influential observations and/or outliers a down-scaling can be applied and the new LDF determined.

## 5.5 USING THE SINGULAR VALUE DECOMPOSITION TO DETERMINE INFLUENTIAL OBSERVATIONS AND/OR OUTLIERS - SIMULTANEOUSLY TAKING CARE OF COLLINEARITY

### 5.5.1 Introduction

In order to apply the SVD to discriminant analysis it is in order to have a look at the relationship between the two subjects.

The application of the basic structure display of a data matrix (BSDM) on the linear discriminant analysis can be described as explained by Greenacre (1980). For the theoretical aspects see appendix B.

Assume n observations in g groups such that $n = n_1 + n_2 + \ldots + n_g$. Then the display which tries to separate the groups maximally can be obtained using BSDM:

$$\text{BSDM}(Z;\ \Omega,\ \phi;\ a,\ b) = \text{BSDM}(\bar{X} - \underline{1}\underline{\bar{x}}';\ \ D_n,\ A^{-1};\ 1,-) \qquad 5.5.1.1$$

where $\bar{X}$: g×m, the matrix of group means on the m variables; $D_n = \text{diag}(n_1,\ldots,n_g)$ and A is the pooled within groups sum of squares and cross products matrix. This provides the co-ordinates of the group centres of gravity in the discriminant subspace. The display of the cases themselves can be obtained using, with $Z = \bar{X} - \underline{1}\underline{\bar{x}}'$:

$$\begin{aligned} F &= (\bar{X} - \underline{1}\underline{\bar{x}}')A^{-\frac{1}{2}}V_1(D_{\mu_1}^{\frac{1}{2}})^{a-1} \\ &= ZA^{-\frac{1}{2}}V_1(D_{\mu_1}^{\frac{1}{2}})^{a-1} \end{aligned} \qquad 5.5.1.2$$

where a = 1, so that

$$F = ZA^{-\frac{1}{2}}V_1 \qquad 5.5.1.3$$

where $V_1$ is the appropriate set of right basic vectors from 5.5.1.1, i.e. the eigenvectors of

$$Q = A^{-\frac{1}{2}}(\bar{X} - \underline{1}\bar{x}')D_n(\bar{X} - \underline{1}\bar{x}')A^{-\frac{1}{2}} \qquad 5.5.1.4$$

Let $\underline{Y} = X\underline{\beta} + \underline{\varepsilon}$ $\quad X: n \times p$

$\beta: (p \times 1)$

$\varepsilon_i \sim n(0, \sigma^2)$ 5.5.1.5

In a similar way to the procedure described in chapter four let the data matrix X take over the role of J, i.e. the jacobian, or the matrix of first derivatives of the model function with respect to the parameters, i.e. exactly what the data matrix is in the linear model as described above.

The data matrix will be standardised as before. Therefore to prevent confusion with too many notational symbolics the X will from hereon represent the standardized data matrix.

As was seen in section 5.4 the detection of outliers or influential observations in discriminant analysis by means of the influence function can be cumbersome. I propose some new approaches as using the SVD of the total data matrix or the group data matrices in order to determine the presence of such values or

observations. The same approach that was used in chapter four will be applied here, i.e. a semi principal components analysis emerges and the abberant values are picked up, removed and the discriminant analysis carried out. During this process the presence of possible collinearities which may emerge can be taken care of.

### 5.5.2 Influence, outliers and collinearity - removing rows and columns

Let

$$X = UD_{a}V', \quad F = UD_{a}, \quad G = VD_{a} \qquad 5.5.2.1$$

From F find those rows that make large contributions to the total variance inherent in X in terms of FF′, i.e. if the total variance is tr(FF′), then

$$tr(FF') = \sum_{i=1}^{n}\left[\sum_{k=1}^{p} f_{ik}^2\right]$$

$$= \sum_{k=1}^{p}\left[\sum_{i=1}^{n} f_{ik}^2\right] \qquad 5.5.2.2$$

where $\sum_{k=1}^{p} f_{ik}^2$ is the contribution of observation i to the total variance. Therefore, the percentage contribution to total variance is expressed as

$$\frac{\sum_{k=1}^{p} f_{ik}^2}{\sum_{i=1}^{n}\left[\sum_{k=1}^{p} f_{ik}^2\right]} \qquad 5.5.2.3$$

We determine therefore, from F those rows that make large contributions to the variance in X. Delete these observations and repeat the analysis. Compare the results, e.g. the misclassification error rate, the discriminant function(s) and so forth with those obtained from the analysis as applied on the original data matrix.

In addition to this determine those columns that make a small contribution through G. As the total variance in X is equal to $tr(GG')$,

$$tr(GG') = \sum_{j=1}^{p} \left[ \sum_{k=1}^{p} g_{jk}^2 \right]$$

$$= \sum_{k=1}^{p} \left[ \sum_{j=1}^{p} g_{jk}^2 \right] \qquad 5.5.2.4$$

where $g_{jk}^2$ represents the variance due to variable k in principal axis j. Then

$$\frac{\sum_{j=1}^{p} g_{jk}^2}{\sum_{j=1}^{p} \sum_{k=1}^{p} g_{jk}^2} \text{ for } k = 1,\ldots,p \qquad 5.5.2.5$$

represents the proportion of the total variance due to variable k.

Determine now those variables with small contributions to the variance, thereby reducing collinearity. Deleting such a variable, or more than one if necessary, create a new data matrix $X_{(i)}$, where (i) refer to the variable(s) left out of the procedure, then

$$X_{(i)} = U_{(i)}D_{a(i)}V'_{(i)}, \quad F_{(i)} = U_{(i)}D_{a(i)}, \quad G_{(i)} = V_{(i)}D_{a(i)} \qquad 5.5.2.6$$

The procedure as described from 5.5.2.2 to the end of 5.5.2.3 and what follows on that may now be repeated in order to determine whether there are still influential observations as such. In fact, the procedure can with care be used in an iterative way, i.e. carry on to 5.5.2.6 and repeat until the procedure stabilizes in the sense that no more collinear variable(s) are indicated or any influential observations come to the fore.

This whole procedure can be approached in a more subtle way by looking at the group data matrices.

If $X = UD_aV'$, $F = UD_a$, $F = VD_a$ then the group data matrices will enter as

$$\begin{aligned} X_1 &= U_1D_{a1}V'_1, \quad F_1 = U_1D_{a1}, \quad G_1 = V_1D_{a1}, \quad X_1: n_1 \times p \\ X_2 &= U_2D_{a2}V'_2, \quad F_2 = U_2D_{a2}, \quad G_2 = V_2D_{a2}, \quad X_2: n_2 \times p, \quad n_1+n_2 = n \end{aligned} \qquad 5.5.2.7$$

Determine in each of these group data matrices those rows that make large contributions to the variances in $X_1$ and $X_2$ respectively, i.e. if

$$\begin{aligned} \mathrm{tr}(F_1F'_1) &= \sum_{i=1}^{n_1}\left[\sum_{k=1}^{p} f^2_{1ik}\right] \\ &= \sum_{k=1}^{p}\left[\sum_{i=1}^{n_1} f^2_{1ik}\right] \end{aligned} \qquad 5.5.2.8$$

where $\sum_{i=1}^{p} f^2_{1ik}$ is the contribution of observation i to the total variance in $X_1$,

then the percentage contribution of observation i to this total group variance is

$$\frac{\sum_{k=1}^{p} f_{1ik}^2}{\sum_{i=1}^{n_1}\left[\sum_{k=1}^{p} f_{1ik}^2\right]} \qquad 5.5.2.9$$

Determine therefore from $F_1$ those rows that make large contributions to the variance in $X_1$. Delete these observations.

Do the same in the case of the second group data matrix $X_2$ (and further if there are more than 2 groups).

Having deleted these group influential observations, put these data matrices together in an adapted total data matrix and repeat the analysis. The results, like the misclassification error rate, the discriminant function(s) etc. can now again be compared to the results of the analysis on the original data matrix.

Once again one can go further and determine the collinear variables in each group. Care however must be taken that corresponding variables are removed in the two groups as one variable may be collinear in one group, but not necessarily in another group. Here it will make sense to have a look at the total data matrix analysis for collinearity as a control as described in 5.5.2.4 up to and including 5.5.2.6. Determine from

$$tr(G_h G_h') = \sum_{j=1}^{p}\left[\sum_{k=1}^{p} g_{hjk}^2\right], \quad h = 1,2$$

$$= \sum_{k=1}^{p} \left[ \sum_{j=1}^{p} g_{hjk}^2 \right] , \qquad 5.5.2.10$$

where $g_{hjk}^2$ represents the variance due to variable k in principal axis j in group data matrix h, the ratio

$$\frac{\sum_{j=1}^{p} g_{hjk}^2}{\sum_{j=1}^{p} \sum_{k=1}^{p} g_{hjk}^2} \text{ for } k = 1,\ldots,p \qquad 5.5.2.11$$

which represents the proportion of the total variance in $X_h$ due to variable k. Delete the variables that make small contributions to the variance thereby reducing collinearity, but do take the remarks before 5.5.2.10 into account. An iterative approach removing group influential observations and group collinearity as far as is practical may be applied again.



### 5.5.3 Influence and collinearity - down weighting procedure

In section 5.5.2 observations which were influential and/or variables which showed collinear characteristics were removed either from the original data matrix directly, or indirectly by an investigation of the group data matrices. This may not necessarily be the optimum solution, which ever way "optimum" may be defined. In stead of removing an influential observation a less drastic procedure would be to rather down weigh an observation if it is influential according to the SVD technique of 5.5.2. This then is a trimming procedure.

One way of down weighing an observation is to replace all variables in an

observation by the arithmetic means of the variables respectively. Other alternatives are to use e.g. the modes or the median or even other types of trimmed means.

Another trimming procedure which may be useful is not to change the data matrix directly, but rather by means of the SVD, i.e. if

$$X = UD_aV' \qquad 5.5.3.1$$

and row (observation) i is influential in the sense of a large contribution to the total variance replace row i in U by the means of the columnar values respectively, thereby creating a new U, i.e. $U^*$. From this adapted U matrix X may be updated so that

$$X^* = U^*D_aV' \qquad 5.5.3.2$$

The discriminant analysis can take place and if collinearities are suspected the procedure described in section 5.5.2 is still applicable.

This trimming procedure does not have to be applied on X directly. It can be applied on the group data matrices $X_1$ and $X_2$, i.e. the group influential observations can be determined and $U_1$ as well as $U_2$ adapted in order to find the trimmed group data matrices $X_1^*$ and $X_2^*$, i.e.

$$X_1^* = U_1^*D_{a1}V_1', \quad X_2^* = U_2^*D_{a2}V_2' .$$

Using $X_1^*$ and $X_2^*$ the group trimmed adapted total matrix $X_{12}^*$ can replace X. Once

again the procedures for detecting collinearities can be applied and an iterative procedure may be followed as before after which an ordinary discriminant analysis procedure may be followed.

## 5.6 Empirical results

The first example which I use is a well-known data set from Johnson and Wichern (1982) in which continuous ratio data is being used in order to allow a firm to establish for itself whether it is on the way of bankruptcy or not. A bank e.g., can use this type of information to assess the solvency of his clients, provided it can get hold of the necessary data. For completeness sake I include the set here - table 5.1.



| CFTD | NITA | CACL | CANS | SOLV |
|---|---|---|---|---|
| -0.448500 | -0.410600 | 1.086500 | 0.452600 | 1 |
| -0.563300 | -0.311400 | 1.513400 | 0.164200 | 1 |
| 0.064300 | 0.015600 | 1.007700 | 0.397800 | 1 |
| -0.072100 | -0.093000 | 1.454400 | 0.258900 | 1 |
| -0.100200 | -0.091700 | 1.564400 | 0.668300 | 1 |
| -0.142100 | -0.065100 | 0.706600 | 0.279400 | 1 |
| 0.035100 | 0.014700 | 1.504600 | 0.708000 | 1 |
| -0.065300 | -0.056600 | 1.373700 | 0.403200 | 1 |
| 0.072400 | -0.007600 | 1.372300 | 0.336100 | 1 |
| -0.135300 | -0.143300 | 1.419600 | 0.434700 | 1 |
| -0.229800 | -0.296100 | 0.331000 | 0.182400 | 1 |
| 0.071300 | 0.020500 | 1.312400 | 0.249700 | 1 |
| 0.010900 | 0.001100 | 2.149500 | 0.696900 | 1 |

| | | | | |
|---|---|---|---|---|
| -0.277700 | -0.231600 | 1.191800 | 0.660100 | 1 |
| 0.145400 | 0.050000 | 1.876200 | 0.272300 | 1 |
| 0.370300 | 0.109800 | 1.994100 | 0.382800 | 1 |
| -0.075700 | -0.082100 | 1.507700 | 0.421500 | 1 |
| 0.045100 | 0.026300 | 1.675600 | 0.949400 | 1 |
| 0.011500 | -0.003200 | 1.260200 | 0.603800 | 1 |
| 0.122700 | 0.105500 | 1.143400 | 0.165500 | 1 |
| -0.284300 | -0.270300 | 1.272200 | 0.152800 | 1 |
| 0.513500 | 0.100100 | 2.487100 | 0.536800 | 2 |
| 0.076900 | 0.019500 | 2.006900 | 0.530400 | 2 |
| 0.377600 | 0.107500 | 3.265100 | 0.354800 | 2 |
| 0.193300 | 0.047300 | 2.250600 | 0.330900 | 2 |
| 0.324800 | 0.071800 | 4.240100 | 0.627900 | 2 |
| 0.313200 | 0.051100 | 4.450000 | 0.685200 | 2 |
| 0.118400 | 0.049900 | 2.521000 | 0.692500 | 2 |
| -0.017300 | 0.023300 | 2.053800 | 0.348400 | 2 |
| 0.216900 | 0.077900 | 2.348900 | 0.397000 | 2 |
| 0.170300 | 0.069500 | 1.797300 | 0.517400 | 2 |
| 0.146000 | 0.051800 | 2.169200 | 0.550000 | 2 |
| -0.098500 | -0.012300 | 2.502900 | 0.577800 | 2 |
| 0.139800 | -0.031200 | 0.461100 | 0.264300 | 2 |
| 0.137900 | 0.072800 | 2.612300 | 0.515100 | 2 |
| 0.148600 | 0.056400 | 2.234700 | 0.556300 | 2 |
| 0.163300 | 0.048600 | 2.308000 | 0.197800 | 2 |
| 0.290700 | 0.059700 | 1.838100 | 0.378600 | 2 |
| 0.538300 | 0.106400 | 2.329300 | 0.483500 | 2 |
| -0.333000 | -0.085400 | 3.012400 | 0.473000 | 2 |

BANKRUPTCY = -5.69076 - 1.48693NITA + 3.88493CACL

For further investigation I used the Statgraphics version 5 routine as the package is more easily available to everyday users. This means that the LDF was used with all the implicated assumptions. Using all the available independent variables an apparent correct classification of 90.48% and 80% for SOLV1 and SOLV2 respectively were obtained.

Applying the SVD of the data set in order to find the relative variance contributions, observations 1 and 46 are the only observations which stand out as possible outliers and/or influential data points, and that with an 8.8% and 8.4% contribution respectively to the total variation only. Note from the table that I standardized the data before running the routine - see table 5.2:

**BANK DATA - 46 OBSERVATIONS**

| i | $\frac{f_{ij}^2}{\sum_j f_{ij}^2}$ | | | | $\sum_j f_{ij}^2$ | $\frac{\sum_j f_{ij}^2}{\sum\sum_{ij} f_{ij}^2}$ |
|---|---|---|---|---|---|---|
| 1 | .849 | .037 | .040 | .075 | .353 | .088 |
| 2 | .819 | .008 | .025 | .148 | .324 | .081 |
| 3 | .220 | .005 | .055 | .720 | .025 | .006 |
| 4 | .740 | .001 | .250 | .009 | .044 | .011 |
| 5 | .290 | .006 | .692 | .012 | .065 | .016 |
| 6 | .759 | .021 | .126 | .094 | .075 | .019 |
| 7 | .003 | .000 | .769 | .227 | .059 | .015 |
| 8 | .936 | .012 | .000 | .052 | .022 | .005 |
| 9 | .308 | .001 | .393 | .298 | .015 | .004 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 10 | .945 | .012 | .033 | .010 | .053 | .013 |
| 11 | .878 | .042 | .078 | .002 | .256 | .064 |
| 12 | .117 | .037 | .667 | .179 | .032 | .008 |
| 13 | .004 | .006 | .979 | .011 | .050 | .012 |
| 14 | .638 | .024 | .334 | .004 | .169 | .042 |
| 15 | .044 | .066 | .835 | .054 | .021 | .005 |
| 16 | .654 | .008 | .142 | .197 | .045 | .011 |
| 17 | .984 | .000 | .015 | .001 | .024 | .006 |
| 18 | .007 | .001 | .883 | .109 | .179 | .045 |
| 19 | .116 | .000 | .500 | .384 | .036 | .009 |
| 20 | .000 | .091 | .671 | .238 | .078 | .020 |
| 21 | .805 | .007 | .099 | .090 | .207 | .052 |
| 22 | .840 | .096 | .010 | .054 | .086 | .021 |
| 23 | .058 | .038 | .776 | .128 | .008 | .002 |
| 24 | .863 | .001 | .078 | .059 | .081 | .020 |
| 25 | .451 | .021 | .528 | .000 | .014 | .003 |
| 26 | .654 | .000 | .156 | .190 | .159 | .040 |
| 27 | .553 | .002 | .228 | .217 | .191 | .048 |
| 28 | .234 | .007 | .755 | .004 | .056 | .014 |
| 29 | .064 | .695 | .230 | .011 | .009 | .002 |
| 30 | .855 | .040 | .088 | .017 | .018 | .004 |
| 31 | .317 | .013 | .142 | .529 | .017 | .004 |
| 32 | .428 | .016 | .452 | .104 | .016 | .004 |
| 33 | .012 | .162 | .716 | .110 | .032 | .008 |
| 34 | .258 | .043 | .327 | .372 | .072 | .018 |
| 35 | .695 | .108 | .194 | .003 | .022 | .006 |
| 36 | .465 | .020 | .443 | .073 | .019 | .005 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 37 | .089 | .052 | .835 | .024 | .040 | .010 |
| 38 | .445 | .037 | .237 | .281 | .021 | .005 |
| 39 | .808 | .103 | .004 | .085 | .086 | .022 |
| 40 | .179 | .199 | .155 | .467 | .091 | .023 |
| 41 | .128 | .050 | .594 | .228 | .111 | .028 |
| 42 | .858 | .013 | .001 | .127 | .203 | .051 |
| 43 | .288 | .052 | .571 | .088 | .024 | .006 |
| 44 | .980 | .007 | .009 | .003 | .095 | .024 |
| 45 | .050 | .032 | .846 | .072 | .060 | .015 |
| 46 | .461 | .014 | .171 | .354 | .336 | .084 |

table 5.2

**BANK DATA - 46 OBSERVATIONS**

| | | | $\frac{f_{ij}^2}{\sum_i f_{ij}^2}$ | |
|---|---|---|---|---|
| 1 | .129950 | .107561 | .014036 | .046565 |
| 2 | .115238 | .022019 | .007964 | .084617 |
| 3 | .002357 | .001053 | .001345 | .031343 |
| 4 | .014248 | .000426 | .011034 | .000700 |
| 5 | .008216 | .003318 | .044890 | .001363 |
| 6 | .024755 | .012905 | .009389 | .012513 |
| 7 | .000086 | .000026 | .044984 | .023664 |
| 8 | .008877 | .002181 | .000002 | .001996 |
| 9 | .001957 | .000146 | .005716 | .007721 |
| 10 | .021575 | .005187 | .001735 | .000947 |

| | | | | |
|---|---|---|---|---|
| 11 | .097500 | .089819 | .019930 | .000848 |
| 12 | .001621 | .009795 | .021115 | .010078 |
| 13 | .000091 | .002338 | .048319 | .000949 |
| 14 | .046874 | .033417 | .056122 | .001069 |
| 15 | .000397 | .011320 | .017085 | .001971 |
| 16 | .012888 | .002898 | .006386 | .015819 |
| 17 | .010166 | .000021 | .000357 | .000035 |
| 18 | .000581 | .001727 | .157038 | .034439 |
| 19 | .001810 | .000067 | .017796 | .024350 |
| 20 | .000006 | .059514 | .052080 | .032863 |
| 21 | .072134 | .011258 | .020326 | .032677 |
| 22 | .031290 | .068672 | .000826 | .008175 |
| 23 | .000211 | .002614 | .006417 | .001888 |
| 24 | .030386 | .000511 | .006252 | .008394 |
| 25 | .002718 | .002440 | .007270 | .000000 |
| 26 | .045135 | .000176 | .024607 | .053391 |
| 27 | .045938 | .002474 | .043248 | .073443 |
| 28 | .005652 | .003024 | .041633 | .000415 |
| 29 | .000254 | .053033 | .002099 | .000184 |
| 30 | .006655 | .005929 | .001565 | .000536 |
| 31 | .002351 | .001803 | .002404 | .015993 |
| 32 | .003030 | .002105 | .007320 | .003002 |
| 33 | .000173 | .043205 | .022822 | .006238 |
| 34 | .008040 | .025557 | .023320 | .047121 |
| 35 | .006780 | .020227 | .004321 | .000139 |
| 36 | .003779 | .003083 | .008225 | .002401 |
| 37 | .001536 | .017373 | .033044 | .001686 |

| | | | | |
|---|---|---|---|---|
| 38 | .004042 | .006476 | .004935 | .010383 |
| 39 | .030292 | .073927 | .000343 | .012982 |
| 40 | .007078 | .150789 | .014055 | .075386 |
| 41 | .006184 | .046600 | .065519 | .044802 |
| 42 | .075414 | .022210 | .000270 | .045423 |
| 43 | .003004 | .010448 | .013621 | .003747 |
| 44 | .040268 | .005686 | .000880 | .000521 |
| 45 | .001309 | .015706 | .050327 | .007605 |
| 46 | .067152 | .038939 | .057029 | .209617 |
| $\sum_i f_{ij}^2$ | 2.3055150 | .1203368 | 1.0080610 | .5660865 |
| $\frac{\sum_i f_{ij}^2}{\sum\sum_{ij} f_{ij}^2}$ | .5763789 | .0300842 | .2520153 | .1415217 |

table 5.3

In the next run I removed observations 1 and 46, but no real mentionable results in terms of a change in the apparent correct classification rate or the discriminant function coefficients resulted.

As the sums of squares of the column values of G and the corresponding percentage of variance due to the respective columns indicated a possibility of redundant variables - see table 5.4.

<u>BANK DATA - 46 OBSERVATIONS</u>

sums of squares of column values of G

2.3055 .1203 1.0081 .5661

column square totals/total variance

.5764 .0301 .2520 .1415

table 5.4

I left out - in addition to observations 1 and 46 - the variables which correspond to columns 2, 3 and 4, i.e. NITA, CACL and CANS. However, as the condition number for the standardized data set was 4.377 only, i.e. much less than 10p = 40 I found that I have lost a lot of information. Accordingly the apparent correct classification rates for SOLV = 1 and SOLV = 2 went down to 85.71% and 84% respectively.

The next step was to determine the relative variance contributions using the SVD on the class group data sets separately and the following results were obtained for SOLV = 1 and SOLV = 2 as separate data sets, containing 21 and 25 observations respectively:

Results for data set one, i.e. SOLV = 1

BANK DATA - 21 OBSERVATIONS - GROUP 1

| i | | $\frac{f_{ij}^2}{\sum_j f_{ij}^2}$ | | | $\sum_j f_{ij}^2$ | $\frac{\sum_j f_{ij}^2}{\sum\sum_{ij} f_{ij}^2}$ |
|---|---|---|---|---|---|---|
| 1 | .816 | .014 | .170 | .000 | .446 | .112 |
| 2 | .664 | .029 | .038 | .270 | .478 | .119 |
| 3 | .071 | .003 | .469 | .457 | .083 | .021 |
| 4 | .052 | .002 | .355 | .591 | .030 | .008 |
| 5 | .118 | .000 | .856 | .026 | .077 | .019 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 6 | .355 | .046 | .346 | .253 | .160 | .040 |
| 7 | .601 | .003 | .228 | .167 | .126 | .032 |
| 8 | .254 | .349 | .378 | .018 | .002 | .000 |
| 9 | .408 | .010 | .581 | .001 | .043 | .011 |
| 10 | .440 | .026 | .412 | .122 | .015 | .004 |
| 11 | .759 | .018 | .115 | .108 | .525 | .131 |
| 12 | .174 | .002 | .818 | .006 | .079 | .020 |
| 13 | .571 | .001 | .348 | .081 | .289 | .072 |
| 14 | .339 | .001 | .552 | .108 | .172 | .043 |
| 15 | .589 | .002 | .126 | .283 | .195 | .049 |
| 16 | .838 | .035 | .061 | .066 | .427 | .107 |
| 17 | .157 | .001 | .188 | .653 | .006 | .002 |
| 18 | .459 | .001 | .417 | .123 | .363 | .091 |
| 19 | .379 | .010 | .044 | .566 | .060 | .015 |
| 20 | .123 | .026 | .844 | .006 | .209 | .052 |
| 21 | .798 | .008 | .015 | .180 | .215 | .054 |

table 5.4.

**BANK DATA - 21 OBSERVATIONS - GROUP 1**

| | | $\frac{f_{ij}^2}{\sum_i f_{ij}^2}$ | | |
|---|---|---|---|---|
| 1 | .154886 | .097278 | .073659 | .000224 |
| 2 | .134800 | .219451 | .017592 | .232050 |
| 3 | .002490 | .003448 | .037608 | .068059 |
| 4 | .000662 | .000856 | .010345 | .031967 |

column square totals/total variance

.5877 .0157 .2579 .1388

table 5.6

Results for data set two, i.e. SOLV = 2

BANK DATA - 25 OBSERVATIONS - GROUP 2

| i | | $\frac{f_{ij}^2}{\sum_j f_{ij}^2}$ | | | $\sum_j f_{ij}^2$ | $\frac{\sum_j f_{ij}^2}{\sum\sum_{ij} f_{ij}^2}$ |
|---|---|---|---|---|---|---|
| 1 | .662 | .083 | .016 | .239 | .124 | .031 |
| 2 | .786 | .024 | .087 | .103 | .076 | .019 |
| 3 | .864 | .092 | .018 | .025 | .093 | .023 |
| 4 | .212 | .021 | .736 | .032 | .022 | .005 |
| 5 | .241 | .000 | .725 | .034 | .185 | .046 |
| 6 | .102 | .016 | .833 | .049 | .249 | .062 |
| 7 | .089 | .002 | .666 | .243 | .125 | .031 |
| 8 | .833 | .076 | .082 | .010 | .096 | .024 |
| 9 | .087 | .386 | .272 | .255 | .013 | .003 |
| 10 | .086 | .012 | .005 | .896 | .046 | .011 |
| 11 | .266 | .000 | .250 | .484 | .039 | .010 |
| 12 | .772 | .000 | .227 | .001 | .217 | .054 |
| 13 | .531 | .116 | .347 | .007 | .365 | .091 |
| 14 | .010 | .290 | .413 | .287 | .027 | .007 |
| 15 | .178 | .002 | .314 | .506 | .039 | .010 |
| 16 | .059 | .080 | .705 | .156 | .092 | .023 |
| 17 | .001 | .058 | .674 | .268 | .029 | .007 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 18 | .710 | .063 | .008 | .218 | .136 | .034 |
| 19 | .789 | .000 | .079 | .132 | .650 | .162 |
| 20 | .116 | .011 | .825 | .048 | .241 | .060 |
| 21 | .868 | .000 | .065 | .067 | .265 | .066 |
| 22 | .001 | .191 | .756 | .051 | .051 | .013 |
| 23 | .892 | .026 | .000 | .081 | .179 | .045 |
| 24 | .063 | .035 | .612 | .291 | .146 | .036 |
| 25 | .262 | .018 | .019 | .701 | .496 | .124 |

table 5.7

**BANK DATA - 25 OBSERVATIONS - GROUP 2**

| | $\frac{f_{ij}^2}{\sum_i f_{ij}^2}$ | | | |
|---|---|---|---|---|
| 1 | .042386 | .075785 | .001706 | .039217 |
| 2 | .030649 | .013360 | .005693 | .010258 |
| 3 | .041572 | .062951 | .001451 | .003105 |
| 4 | .002381 | .003297 | .013842 | .000925 |
| 5 | .022856 | .000371 | .115266 | .008273 |
| 6 | .013084 | .028868 | .178874 | .016155 |
| 7 | .005738 | .001747 | .071882 | .040240 |
| 8 | .040970 | .052753 | .006713 | .001306 |
| 9 | .000604 | .037988 | .003149 | .004533 |
| 10 | .002040 | .004134 | .000207 | .054314 |
| 11 | .005305 | .000023 | .008320 | .024705 |
| 12 | .086066 | .000130 | .042271 | .000234 |

| | | | | |
|---|---|---|---|---|
| 13 | .099672 | .308820 | .109015 | .003159 |
| 14 | .000139 | .056242 | .009438 | .010047 |
| 15 | .003544 | .000675 | .010434 | .025778 |
| 16 | .002782 | .053556 | .055669 | .018949 |
| 17 | .000008 | .012375 | .017114 | .010413 |
| 18 | .049793 | .062533 | .000972 | .039249 |
| 19 | .263827 | .000468 | .044386 | .112935 |
| 20 | .014390 | .018901 | .170991 | .015388 |
| 21 | .118430 | .000123 | .014750 | .023382 |
| 22 | .000032 | .070463 | .032884 | .003421 |
| 23 | .082282 | .034079 | .000028 | .019266 |
| 24 | .004685 | .036893 | .076660 | .055823 |
| 25 | .066768 | .063465 | .008284 | .458925 |
| $\sum_i f_{ij}^2$ | 1.9435450 | .1370284 | 1.1616430 | .7577857 |
| $\frac{\sum_i f_{ij}^2}{\sum\sum_{ij} f_{ij}^2}$ | .4858860 | .0342570 | .2904105 | .1894463 |

table 5.8



**BANK DATA - 25 OBSERVATIONS - GROUP 2**

sums of squares of column values of G

1.9435 .1370 1.1616 .7578

column squares totals/total variance

.4859 .0343 .2904 .1894

table 5.9

If we look at table 5.4 then observations 1, 2, 11 and 16 stand out as possible outliers/influential observations with relative variance contributions of 11.2%, 11.9%, 13.1% and 10.7% respectively. The next observation in terms of relative variance contribution is observation number 18 with 9.1%.

As the condition numbers are fairly small i.e. 6.12 and 3.77, I did not think of removing any variable in spite of the columnar relative variance contributions as indicated in tables 5.6 and 5.9.

In table 5.7 the possible outliers/influential observations are observations number 19 and 25 with relative variance contributions of 16.2% and 12.4% respectively, i.e. in the basic data set observations 21 + 19 = 40 and 21 + 25 = 46 respectively.

The last run was made, leaving out observations 1, 2, 11, 16, 40 (19 in the second data set) and 46 (25 in the second data set).

The apparent correct classification rate now changed to 90.48 for SOLV = 1 and 92% for SOLV = 2, a rather significant increase in accuracy.

This whole procedure could now be repeated by looking at the adapted data sets 1 and 2 and doing the separate SVD relative variance contribution analysis again, and then one could try to determine further possible outliers/influential observations.

It is interesting to see how possible outliers/influential observations are masked in the full data set SVD, except maybe for observations 1, 2 and 46. When

each group is analized separately however, the smaller covariance matrices (convex hulls) result in pointing out possible deviant observations.

I ran the same procedures for the centered data sets and found interestingly enough that there is a marked correspondence in the results, but in a sense also a marked difference. The SVD at the basic data set pointed out that observations 11, 26, 27, 42 and 46 have variance contributions of 6.2%, 9.8%, 11.7%, 10.2% and 18.6% respectively.

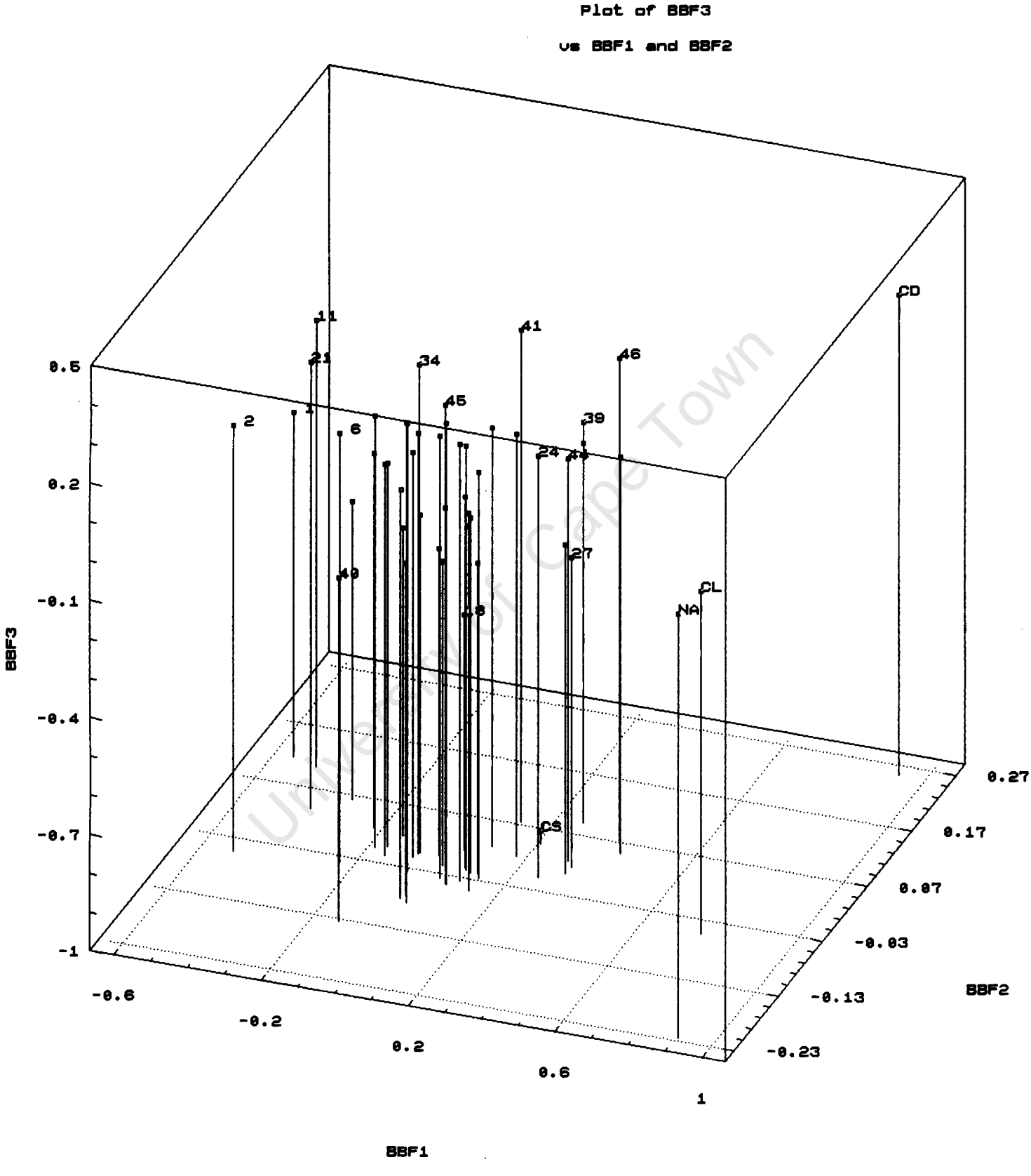
In the case of groups 1 and 2 separately the following observations were identified as possible outliers/influential: 11 (21.7%), 16 (11.2%), 27 (13.1%), 34 (17.1%), 40 (11.1%) and 46 (23.4%).

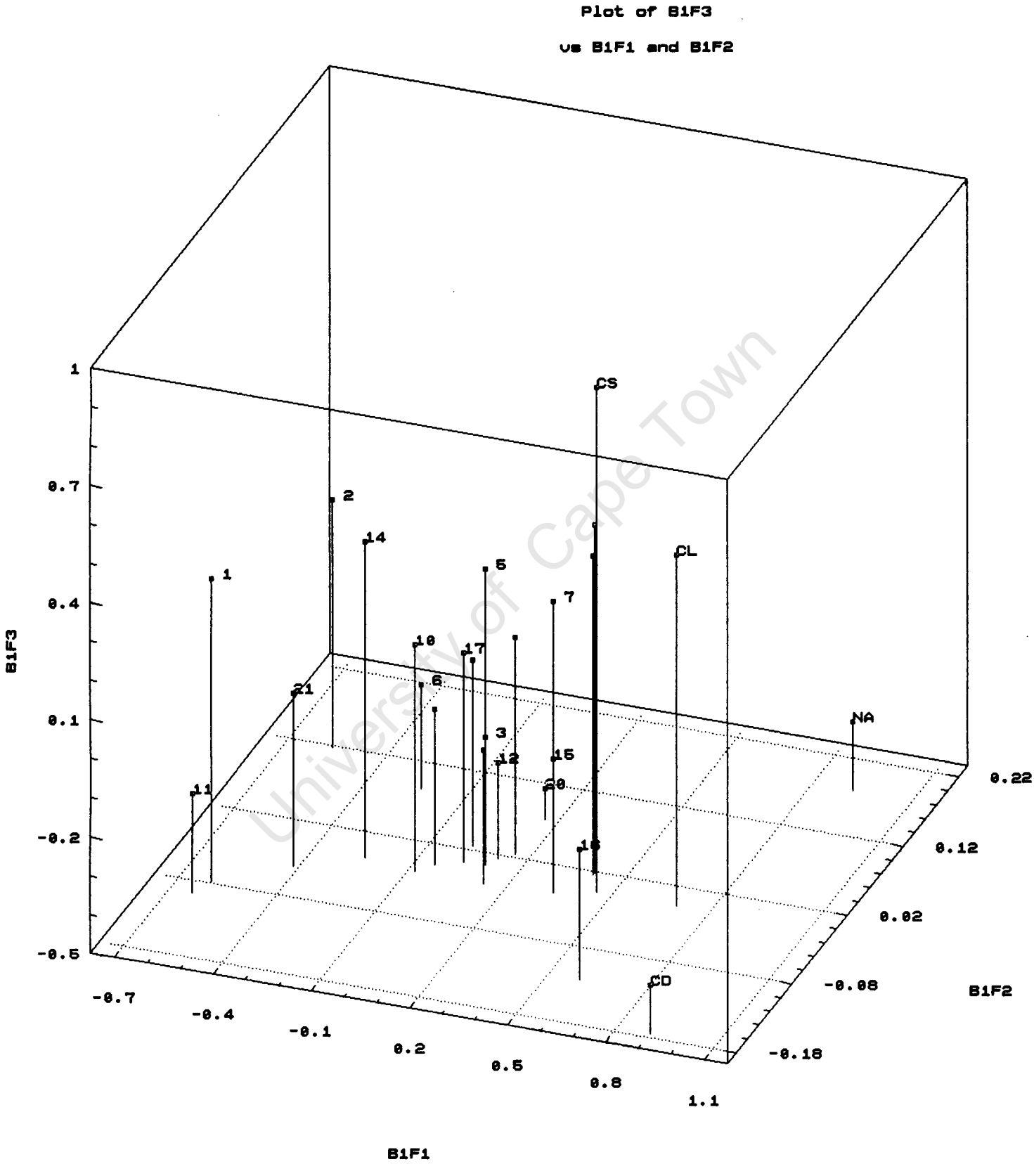
If one compares the means and standard deviations of the variables in groups 1 and 2 one finds that the means of variables 1 and 3 are significantly different (assuming normal distribution universa) with the standard deviations probably due for further investigation - see later.

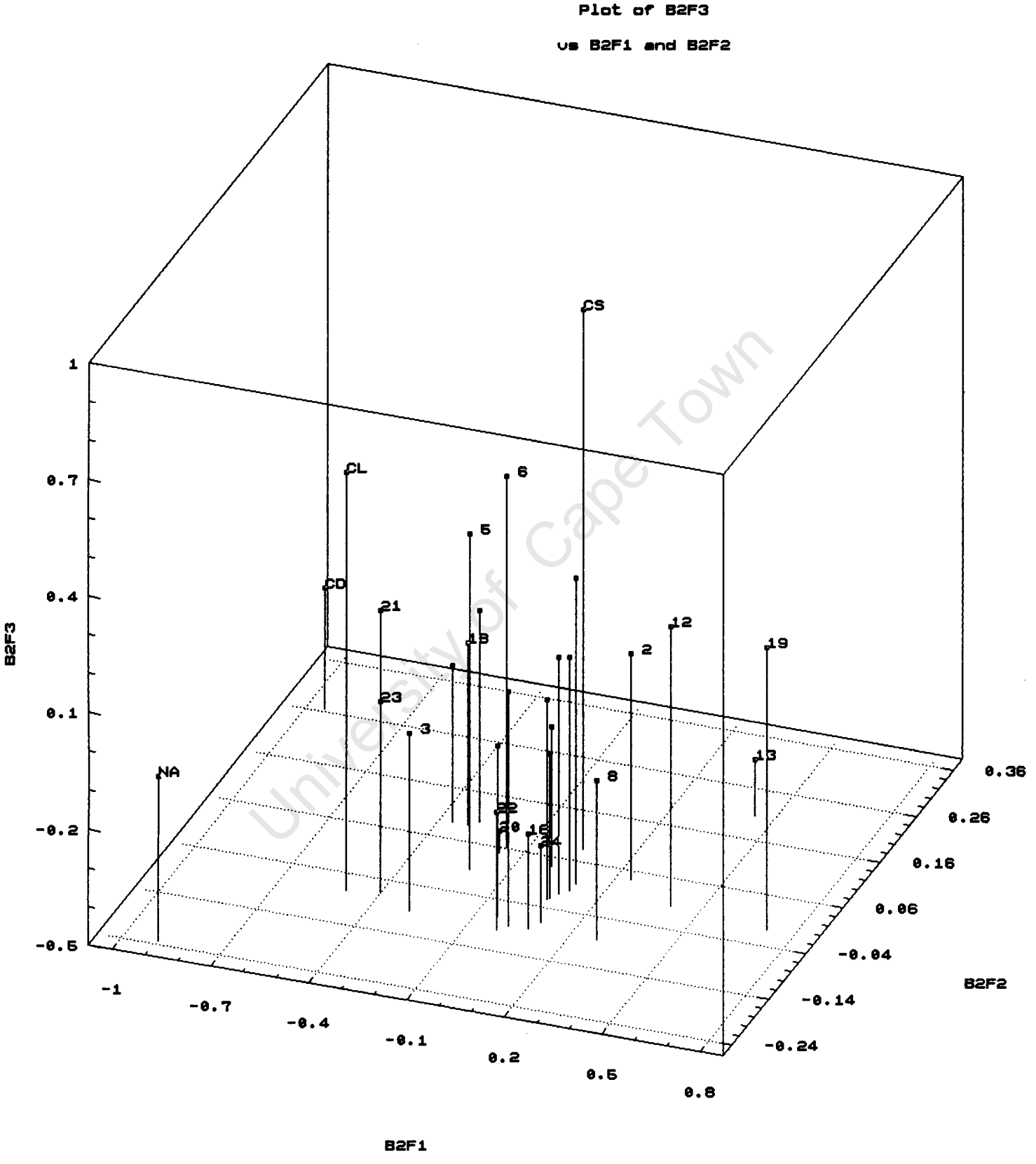
As a further tool of investigation I used the 3-dimensional biplots of the SVD with F against G. The Statgraphics routine was applied. Note that the abbreviations are: -

BBFi ≡ ith principal component of basic bank data.
B1Fi ≡ ith principal component of bank data group 1.
B2Fi ≡ ith principal component of bank data group 2.

The graphical displays are on the next few pages.


Plot of BBF3
vs BBF1 and BBF2
BBF3
BBF1
BBF2
0.5
0.2
-0.1
-0.4
-0.7
-1
-0.6
-0.2
0.2
0.6
1
0.27
0.17
0.07
-0.03
-0.13
-0.23
11
21
34
41
46
45
39
2
6
24
40
27
CD
CL
NA
CS
University of Cape Town





Observing these graphical displays refer to the last rows of tables 5.3, 5.5 and 5.8 from which it is clear that principal component 1 is responsible for most of the variance in all cases. This means that the weights of the other two principal components in each graph are somewhat over-emphasized by the displays. (Full set: 57%; group 1: 59%; group 2: 49%).

BBF1 vs BBF2 vs BBF3: Observations 1, 2, 11 and 46 are all four on the rim of the convex hull with respect to BBF1. With respect to BBF2 observation 40 is on the rim of the hull and so is observation 41 and 46 with respect to principal component 3. The unmarked observation just below observation 46 is observation nr 42 - a rather surprising result with relative variance contributions of 5.1% (full data set) and 6.6% (group 2 data set).

B1F1 vs B1F2 vs B1F3: Here we find observations 1, 2 and 11 on the far left hand side of the convex hull with observation 16 on the far right hand side (slightly obscured by column CS).

B2F1 vs B2F2 vs B2F3: Here we find observation 19 (i.e. 21 + 19 = 40) on the right hand side rim of the convex hull with observation 21 (i.e. 21 + 21 = 42) on the left hand side. Observation 42 seems to be an interesting point - the reason for its graphical placement is not clear to me. Observation 25 is just to the left hand side of observation 18 which is on the rim of the convex hull defined by principal component 1 together with principal component 2. Taking all this into account leaving out the observations as quoted in the second run earlier had the effect shown and the apparent correct classification rate because of their influence as indicated by the analyses that followed.

Eventually some more remarks and comparisons will be shown with respect to this data set, especially with regard to other discriminant analysis techniques - some of them quite robust in approach.

The next example entails a data set which was obtained by means of a type of stratified sampling method. (See Jacobs, 1983 and Van Deventer, 1985). The Republic of South Africa was divided into 13 regions and the common usage by the public of the public media was investigated with the aim of investigating the optimal use of media options for advertising purposes - depending on the market for which one wanted to cater. The question was: If one has a certain population (market) in mind for advertising purposes, can one discriminate correctly by paying advertising fees only to certain types of papers, radio stations etc.? Stated differently: Can one make any remarks with respect to a company's efficiency in the sense that the right market is being reached ? The data set and its analysis is interesting not only in terms of the good discrimination that evolved, but also in terms of the correspondence between the good discrimination and the almost complete absence of influential observations/outliers. The investigation of the covariance matrices however lead to some interesting side-results - see later.



The full tabular display of the analysis, i.e. the SVD, principal components and variance contributions etc are not supplied as the full Media data set consists of 4 classifications and 92 observations. Only the data set is supplied as well as the graphical displays of 3-dimensional biplots of some of the principal components together with the variable principal components.

Legend: -

1. CIN = % who attended a cinema showing the last two weeks
2. ENGDAY = % reading English daily newspaper
3. AFRDAY = % reading Afrikaans daily newspaper
4. ANYW = % reading any weekly newspaper
5. MAGA = % reading any magazine
6. TELV = % watching prime television (20:00 - 21:30)
7. RADV = % listening to Radio 5
8. SPRI = % listening to Springbok Radio (now a regional station)
9. CLAGRO = class group:
    1. EURO = Europeans ($\pi_1$)
    2. COLOU = Colourdes ($\pi_2$)
    3. AFRI = Africans ($\pi_3$)
    4. ASIAN = Asian ($\pi_4$)



ADVERTISING MEDIA

| | CIN | ENGD | AFRD | ANYW | MAGA | TELV | RADV | SPRI | CLAGRO | |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 20.3 | 45.4 | 34.7 | 86.0 | 98.8 | 60.1 | 15.8 | 24.9 | 1 | 1 |
| 2 | 26.4 | 69.8 | 6.3 | 85.8 | 94.9 | 57.8 | 20.1 | 31.8 | 1 | 1 |
| 3 | 23.5 | 40.6 | 33.2 | 79.2 | 91.9 | 56.2 | 17.1 | 27.2 | 1 | 1 |
| 4 | 22.2 | 15.9 | 51.2 | 76.3 | 89.8 | 69.2 | 12.4 | 33.6 | 1 | 1 |
| 5 | 23.9 | 56.4 | 27.3 | 83.2 | 90.9 | 56.9 | 21.0 | 27.7 | 1 | 1 |
| 6 | 24.5 | 26.6 | 37.5 | 78.6 | 91.0 | 62.4 | 11.8 | 29.0 | 1 | 1 |
| 7 | 17.8 | 18.0 | 41.5 | 81.8 | 93.7 | 58.4 | 7.5 | 29.2 | 1 | 1 |
| 8 | 14.5 | 19.4 | 39.0 | 77.3 | 93.7 | 62.2 | 7.6 | 24.1 | 1 | 1 |
| 9 | 15.3 | 42.1 | 33.5 | 81.0 | 91.7 | 64.0 | 11.6 | 28.4 | 1 | 1 |
| 10 | 48.2 | 48.3 | 27.0 | 83.8 | 92.7 | 42.8 | 35.5 | 25.3 | 1 | 1 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 11 | 46.1 | 39.3 | 32.1 | 82.3 | 95.4 | 48.3 | 37.0 | 30.5 | 1 | 1 |
| 12 | 28.3 | 45.5 | 30.7 | 84.0 | 94.2 | 55.1 | 21.2 | 25.7 | 1 | 1 |
| 13 | 12.9 | 48.5 | 34.0 | 83.2 | 92.1 | 63.0 | 8.9 | 23.8 | 1 | 1 |
| 14 | 18.6 | 38.7 | 26.6 | 81.0 | 91.3 | 63.8 | 11.4 | 32.5 | 1 | 1 |
| 15 | 20.4 | 64.8 | 2.7 | 83.5 | 94.4 | 61.5 | 15.5 | 42.3 | 1 | 1 |
| 16 | 23.4 | 34.4 | 25.6 | 74.2 | 93.2 | 63.9 | 11.5 | 33.9 | 1 | 1 |
| 17 | 16.3 | 16.8 | 53.0 | 67.0 | 95.3 | 61.8 | 6.5 | 42.0 | 1 | 1 |
| 18 | 25.2 | 50.6 | 22.0 | 78.0 | 92.5 | 63.5 | 15.6 | 35.8 | 1 | 1 |
| 19 | 20.3 | 17.1 | 28.2 | 75.6 | 91.9 | 64.0 | 7.4 | 32.9 | 1 | 1 |
| 20 | 9.4 | 14.6 | 33.0 | 80.9 | 93.3 | 66.1 | 3.5 | 34.2 | 1 | 1 |
| 21 | 10.1 | 17.6 | 29.2 | 67.9 | 96.5 | 59.3 | 2.6 | 35.3 | 1 | 1 |
| 22 | 16.1 | 36.5 | 26.1 | 77.7 | 93.8 | 67.4 | 8.5 | 35.7 | 1 | 1 |
| 23 | 50.0 | 46.9 | 27.4 | 79.1 | 96.9 | 45.5 | 33.8 | 31.3 | 1 | 1 |
| 24 | 42.5 | 38.9 | 23.9 | 79.2 | 97.1 | 54.4 | 29.4 | 36.8 | 1 | 1 |
| 25 | 22.6 | 38.2 | 27.0 | 77.1 | 95.9 | 63.6 | 13.1 | 32.9 | 1 | 1 |
| 26 | 15.7 | 35.8 | 26.9 | 80.1 | 95.3 | 67.8 | 6.3 | 33.1 | 1 | 1 |
| 27 | 12.8 | 29.1 | 13.0 | 58.8 | 48.7 | 27.1 | 9.8 | 21.6 | 2 | 1 |
| 28 | 19.9 | 55.0 | 5.1 | 67.1 | 48.6 | 51.4 | 7.5 | 41.7 | 2 | 1 |
| 29 | 16.2 | 55.1 | 8.0 | 77.6 | 60.4 | 45.5 | 13.3 | 28.4 | 2 | 1 |
| 30 | 22.0 | 25.8 | 31.3 | 67.1 | 58.8 | 30.5 | 17.5 | 23.7 | 2 | 1 |
| 31 | 13.0 | 4.3 | 16.9 | 51.9 | 43.2 | 13.9 | 4.7 | 21.2 | 2 | 1 |
| 32 | 3.7 | 0.2 | 10.4 | 23.3 | 23.8 | 3.5 | 1.7 | 13.1 | 2 | 1 |
| 33 | 6.2 | 32.8 | 11.5 | 59.3 | 44.2 | 33.0 | 7.2 | 23.9 | 2 | 1 |
| 34 | 26.1 | 29.6 | 15.3 | 63.1 | 59.3 | 23.4 | 14.4 | 23.2 | 2 | 1 |
| 35 | 29.3 | 31.8 | 14.7 | 65.9 | 60.8 | 26.5 | 15.8 | 24.0 | 2 | 1 |
| 36 | 8.4 | 35.5 | 11.5 | 60.5 | 54.2 | 30.2 | 11.7 | 17.9 | 2 | 1 |
| 37 | 5.1 | 32.3 | 15.2 | 62.5 | 44.0 | 32.8 | 5.1 | 26.2 | 2 | 1 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 38 | 7.5 | 26.5 | 10.0 | 57.3 | 60.9 | 28.9 | 9.4 | 31.1 | 2 | 1 |
| 39 | 14.4 | 36.7 | 3.1 | 57.5 | 62.9 | 49.7 | 8.6 | 52.9 | 2 | 1 |
| 40 | 10.1 | 47.6 | 6.2 | 72.1 | 72.7 | 47.0 | 13.9 | 37.5 | 2 | 1 |
| 41 | 6.5 | 13.0 | 19.5 | 63.2 | 69.0 | 31.8 | 7.9 | 40.3 | 2 | 1 |
| 42 | 8.5 | 2.0 | 19.8 | 46.4 | 55.0 | 10.5 | 2.1 | 26.0 | 2 | 1 |
| 43 | 2.7 | 1.5 | 4.0 | 22.8 | 29.9 | 2.0 | 3.8 | 21.4 | 2 | 1 |
| 44 | 4.8 | 28.4 | 8.1 | 55.0 | 58.2 | 36.1 | 5.8 | 34.9 | 2 | 1 |
| 45 | 15.4 | 29.8 | 12.1 | 65.6 | 72.9 | 24.2 | 15.7 | 29.2 | 2 | 1 |
| 46 | 15.1 | 30.2 | 10.8 | 63.7 | 77.5 | 26.5 | 16.8 | 28.3 | 2 | 1 |
| 47 | 8.9 | 28.8 | 12.2 | 58.3 | 64.4 | 33.9 | 6.7 | 29.9 | 2 | 1 |
| 48 | 1.4 | 28.2 | 8.8 | 56.6 | 54.5 | 33.7 | 5.7 | 38.2 | 2 | 1 |
| 49 | 31.1 | 71.9 | 1.2 | 86.2 | 63.7 | 38.2 | 31.3 | 37.3 | 3 | 1 |
| 50 | 41.7 | 69.4 | 1.0 | 78.8 | 60.2 | 50.3 | 14.8 | 24.9 | 3 | 1 |
| 51 | 35.7 | 77.0 | 1.3 | 88.7 | 64.2 | 42.7 | 31.5 | 33.3 | 3 | 1 |
| 52 | 21.6 | 49.4 | 1.1 | 68.5 | 46.7 | 22.4 | 23.8 | 37.4 | 3 | 1 |
| 53 | 18.7 | 71.2 | 0.7 | 85.2 | 56.9 | 42.6 | 21.7 | 35.4 | 3 | 1 |
| 54 | 59.5 | 73.4 | 2.0 | 86.3 | 77.2 | 36.8 | 43.5 | 36.4 | 3 | 1 |
| 55 | 58.1 | 72.2 | 3.1 | 85.7 | 76.0 | 34.5 | 41.9 | 35.0 | 3 | 1 |
| 56 | 28.7 | 73.6 | 0.3 | 90.3 | 70.6 | 39.2 | 35.2 | 42.6 | 3 | 1 |
| 57 | 16.1 | 73.9 | 0.0 | 88.2 | 59.1 | 47.4 | 17.8 | 28.7 | 3 | 1 |
| 58 | 17.9 | 46.6 | 0.0 | 69.5 | 54.2 | 42.9 | 25.9 | 37.9 | 3 | 1 |
| 59 | 32.2 | 43.2 | 0.5 | 65.8 | 59.5 | 51.7 | 13.4 | 38.5 | 3 | 1 |
| 60 | 22.4 | 50.7 | 0.1 | 72.3 | 59.1 | 48.5 | 27.1 | 38.4 | 3 | 1 |
| 61 | 10.7 | 25.6 | 0.0 | 49.0 | 33.4 | 20.7 | 15.5 | 31.7 | 3 | 1 |
| 62 | 16.3 | 45.7 | 0.0 | 70.0 | 53.5 | 46.3 | 18.8 | 42.5 | 3 | 1 |
| 63 | 37.9 | 59.2 | 0.3 | 82.6 | 76.1 | 42.8 | 46.7 | 27.3 | 3 | 1 |
| 64 | 30.7 | 51.8 | 0.2 | 80.9 | 75.7 | 37.8 | 40.6 | 32.2 | 3 | 1 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 65 | 16.6 | 52.8 | 0.0 | 75.4 | 58.4 | 49.2 | 22.1 | 49.1 | 3 | 1 |
| 66 | 13.3 | 43.0 | 0.0 | 63.2 | 46.4 | 51.8 | 12.4 | 39.6 | 3 | 1 |
| 67 | 8.3 | 11.9 | 0.8 | 21.3 | 24.4 | 4.8 | 2.4 | 2.5 | 4 | 1 |
| 68 | 10.5 | 9.9 | 0.2 | 46.6 | 42.0 | 3.5 | 3.1 | 10.6 | 4 | 1 |
| 69 | 21.4 | 17.6 | 1.1 | 22.7 | 35.2 | 3.9 | 2.5 | 1.9 | 4 | 1 |
| 70 | 10.0 | 9.0 | 4.8 | 14.9 | 34.2 | 3.2 | 0.4 | 0.9 | 4 | 1 |
| 71 | 25.3 | 28.0 | 1.7 | 41.8 | 44.7 | 6.5 | 4.1 | 3.3 | 4 | 1 |
| 72 | 35.0 | 18.7 | 2.9 | 29.2 | 47.7 | 6.4 | 2.8 | 3.2 | 4 | 1 |
| 73 | 25.8 | 17.9 | 2.9 | 29.2 | 50.0 | 7.6 | 2.3 | 2.3 | 4 | 1 |
| 74 | 3.8 | 3.8 | 0.3 | 15.9 | 22.6 | 1.2 | 1.3 | 4.1 | 4 | 1 |
| 75 | 8.2 | 13.1 | 1.2 | 21.5 | 24.8 | 3.2 | 1.3 | 2.4 | 4 | 1 |
| 76 | 24.5 | 15.6 | 1.1 | 32.7 | 46.3 | 5.1 | 3.7 | 5.3 | 4 | 1 |
| 77 | 26.0 | 10.1 | 1.4 | 29.6 | 46.9 | 4.0 | 4.4 | 4.6 | 4 | 1 |
| 78 | 22.7 | 25.0 | 1.7 | 37.3 | 47.9 | 6.7 | 2.7 | 4.9 | 4 | 1 |
| 79 | 5.9 | 14.8 | 1.3 | 23.4 | 24.7 | 3.3 | 1.6 | 2.2 | 4 | 1 |
| 80 | 2.6 | 3.6 | 0.5 | 10.7 | 16.1 | 1.7 | 0.6 | 0.8 | 4 | 1 |
| 81 | 3.0 | 4.3 | 0.0 | 33.4 | 33.8 | 0.4 | 2.2 | 0.9 | 4 | 1 |
| 82 | 3.0 | 5.3 | 0.6 | 11.7 | 24.7 | 2.4 | 0.6 | 0.8 | 4 | 1 |
| 83 | 0.3 | 4.0 | 1.5 | 5.4 | 22.2 | 1.8 | 1.2 | 0.2 | 4 | 1 |
| 84 | 7.1 | 17.4 | 1.5 | 34.4 | 44.7 | 5.2 | 2.3 | 2.6 | 4 | 1 |
| 85 | 4.6 | 6.0 | 2.6 | 26.1 | 42.3 | 10.0 | 0.9 | 1.6 | 4 | 1 |
| 86 | 0.8 | 4.5 | 0.0 | 15.2 | 36.1 | 1.7 | 2.5 | 1.2 | 4 | 1 |
| 87 | 1.4 | 1.0 | 0.2 | 10.5 | 16.8 | 0.2 | 0.6 | 0.2 | 4 | 1 |
| 88 | 1.3 | 3.9 | 0.3 | 11.5 | 16.0 | 1.8 | 0.7 | 0.7 | 4 | 1 |
| 89 | 6.3 | 6.3 | 1.1 | 24.7 | 44.3 | 2.2 | 2.1 | 1.1 | 4 | 1 |
| 90 | 6.4 | 4.9 | 0.9 | 20.2 | 42.4 | 1.9 | 2.3 | 1.2 | 4 | 1 |
| 91 | 2.0 | 5.7 | 0.6 | 18.0 | 23.1 | 2.6 | 1.4 | 0.9 | 4 | 1 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 92 | 1.2 | 5.7 | 0.4 | 17.7 | 21.7 | 1.7 | 0.1 | 0.9 | 4 | 1 |

Using the LDF approach in the BMDP7M routines an apparent rate of correct classification of 96.7% was obtained (Apparent Error Rate = 3.3%).

Apply the SVD approach to the full data set as well as each group separately as in the last example.

The following "significant" relative variance contributors could be identified: i.e. an approximate relative variance contribution of $\frac{2}{n} \times 100\%$ and more was arbitrarily chosen as "significant".

Full data set (2.17%): Observations 54 (3.34%), 83 (2.19%), 87 (2.26%) and 88 (2.17%); i.e. not one really of any importance, except observation 54 (maybe), (Cond. no: 13.63).

Class group 1 (7.7%): Observations 10 (9.74%), 17 (9.86%) and 23 (8.38%); (Cond. no: 15.5).

Class group 2 (9.1%): Observations 32 (i.e. no 6 in group 2) (15.7%) and 43 (i.e. no 17 in group 2) (13.8%); (Cond. no: 17).

Class group 3 (11.1%): Observations 55 (i.e. no 7 in group 3) (13.5%), and 61 (i.e. no 13 in group 3) (18.9%); (Cond. no: 63).

Class group 4 (7.7%): Observations 68 (i.e. no 2 in group 4) (10.6%), 71 (i.e. no 5 in group 4) (8.9%); (Cond. no: 8.8).

The relative column (principal component) importance of principal component 1 for the different groups are:

Full data set: Principal Component 1: 65%
Group 1 : Principal Component 1: 46%
Group 2 : Principal Component 1: 42%
Group 3 : Principal Component 1: 53%
Group 4 : Principal Component 1: 61%

The 3-dimensional biplots of F with G are coded as follows:

MjFi: M = Media
j = B ≡ Full data set
= j ≡ group j
F is the principal component of F with G
i = ith principal component

It will be noticed that not all the principal components and certainly not all the combinations are shown, because of the small contributions of many of these components. Keep in mind the relative importance of principal component 1 throughout.

MBF1 vs MBF2 vs MBF3: The discrimination, i.e. in this case, classification is good. Group one is the L.H.S. "column", group 2 is somewhere between the middle "column" and R.H.S. "column" somewhat lower down. Note the position of observations 43 (with respect to the position of the 2nd group observations), 61 with respect to the 3rd group observations and 87 with respect to the full set as

well as with respect to group 4.

MBF1 vs MBF3 vs MBF4: Observation 88 on the rim of the convex hull.

MBF1 vs MBF4 vs MBF8: Observations 55, 61, 87 are at the rim of the convex hull.

MBF2 vs MBF3 vs MBF4: In the absence of the main principal component no classes can be recognized except for group 3 on the R.H.S.

M1F1 vs M1F2 vs M1F3: Observations 10, 17 and 23 stand out as predicted by the SVD as possible influential observations/outliers.

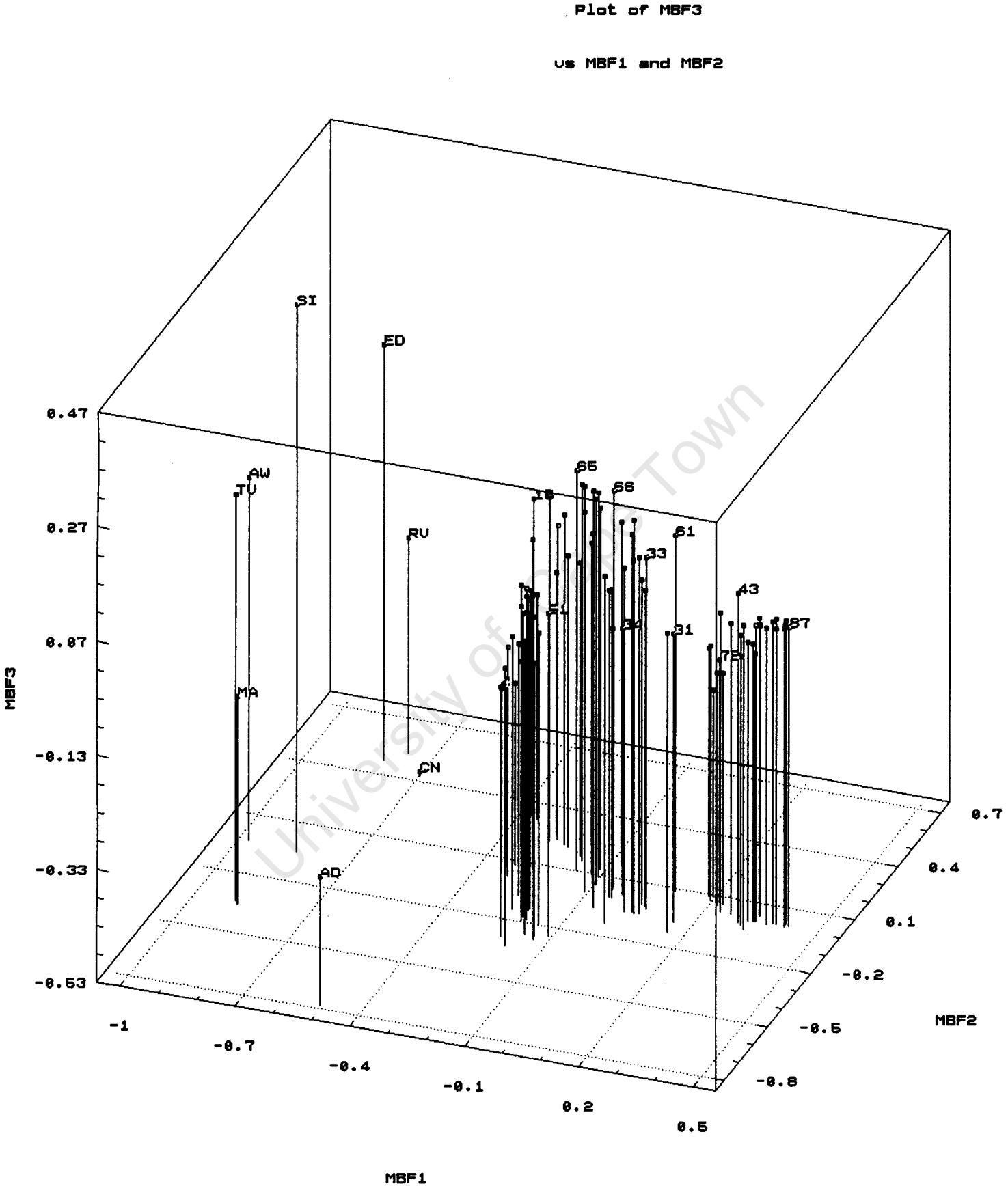
M1F1 vs M1F3 vs M1F4<br>
M1F1 vs M1F4 vs M1F8<br>
M1F2 vs M1F3 vs M1F4

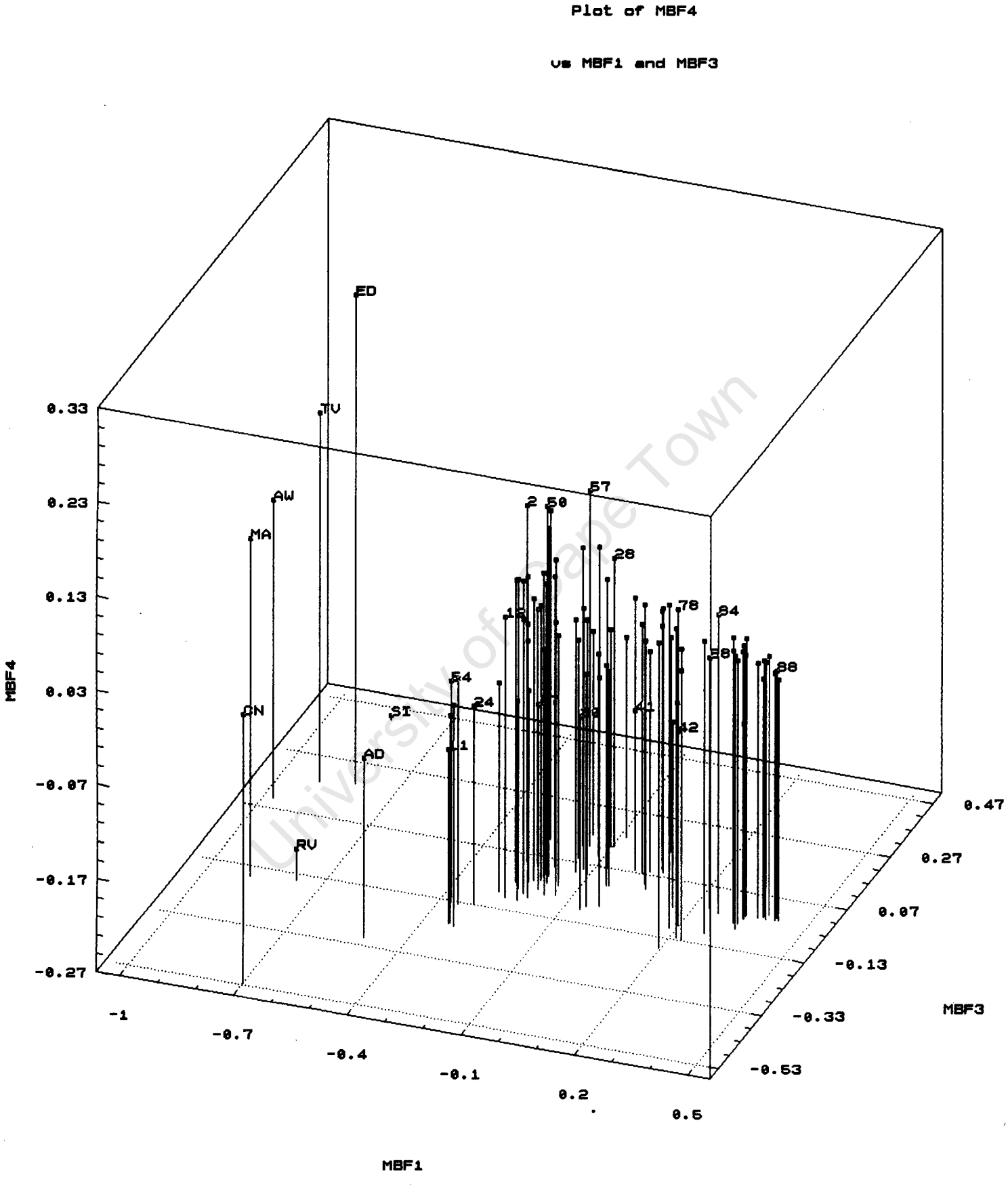
} are self explanatory.

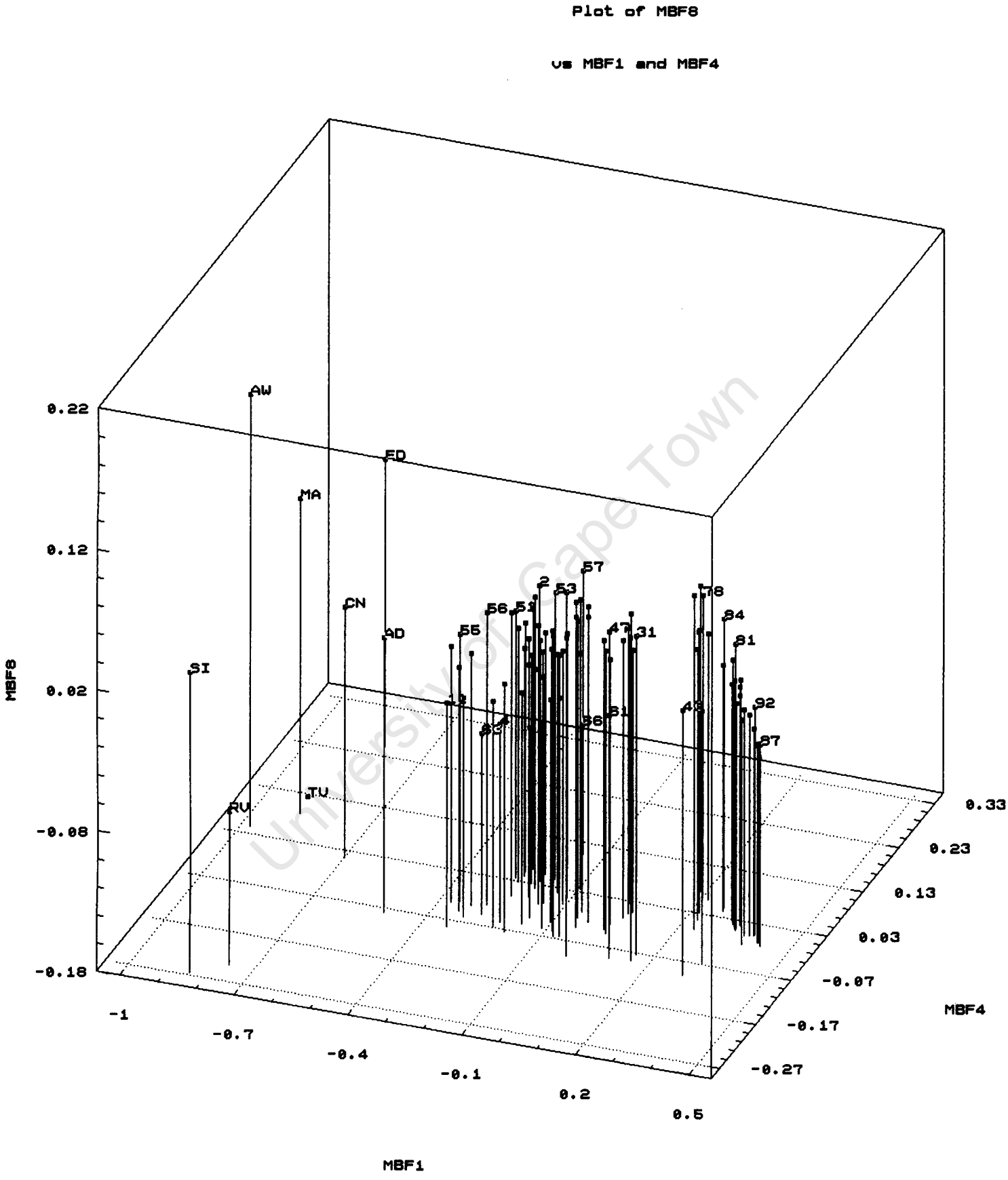
M2F1 vs M2F2 vs M2F3: Observations 32 and 43 stand out as in all the plots of M2F1 versus the others. In M2F2 vs M2F3 vs M2F4 the 1st principal component has no effect, so that no deductions can be made.

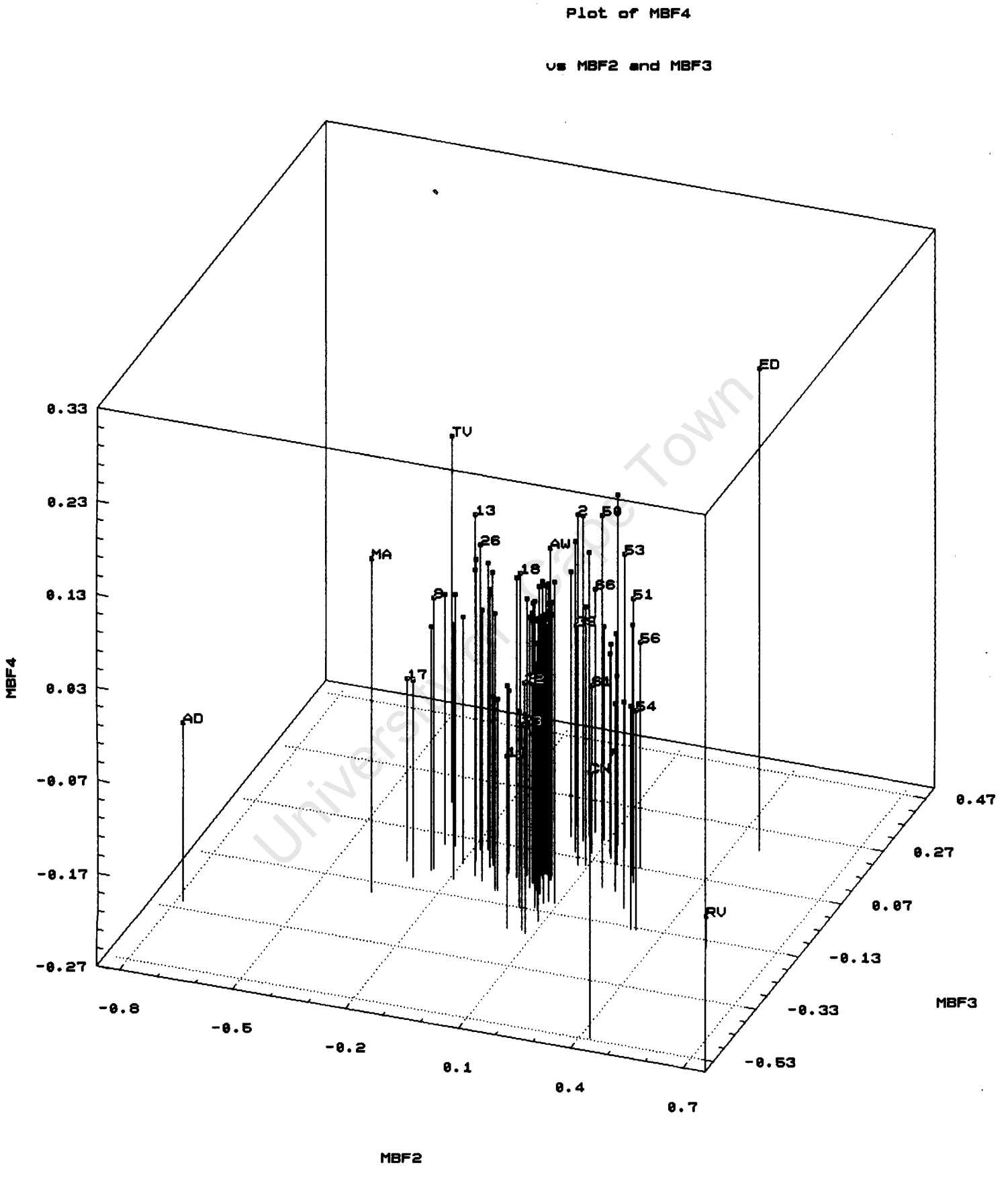
M3F1 vs M3F2 vs M3F3: Observations 55 and 61 are on the outside rim of the convex hull. Considering M3F1 observations 54 and 66 seem as if they want to qualify, but the size of this sample doesn't satisfy the demand for enough information before decisions can be made.

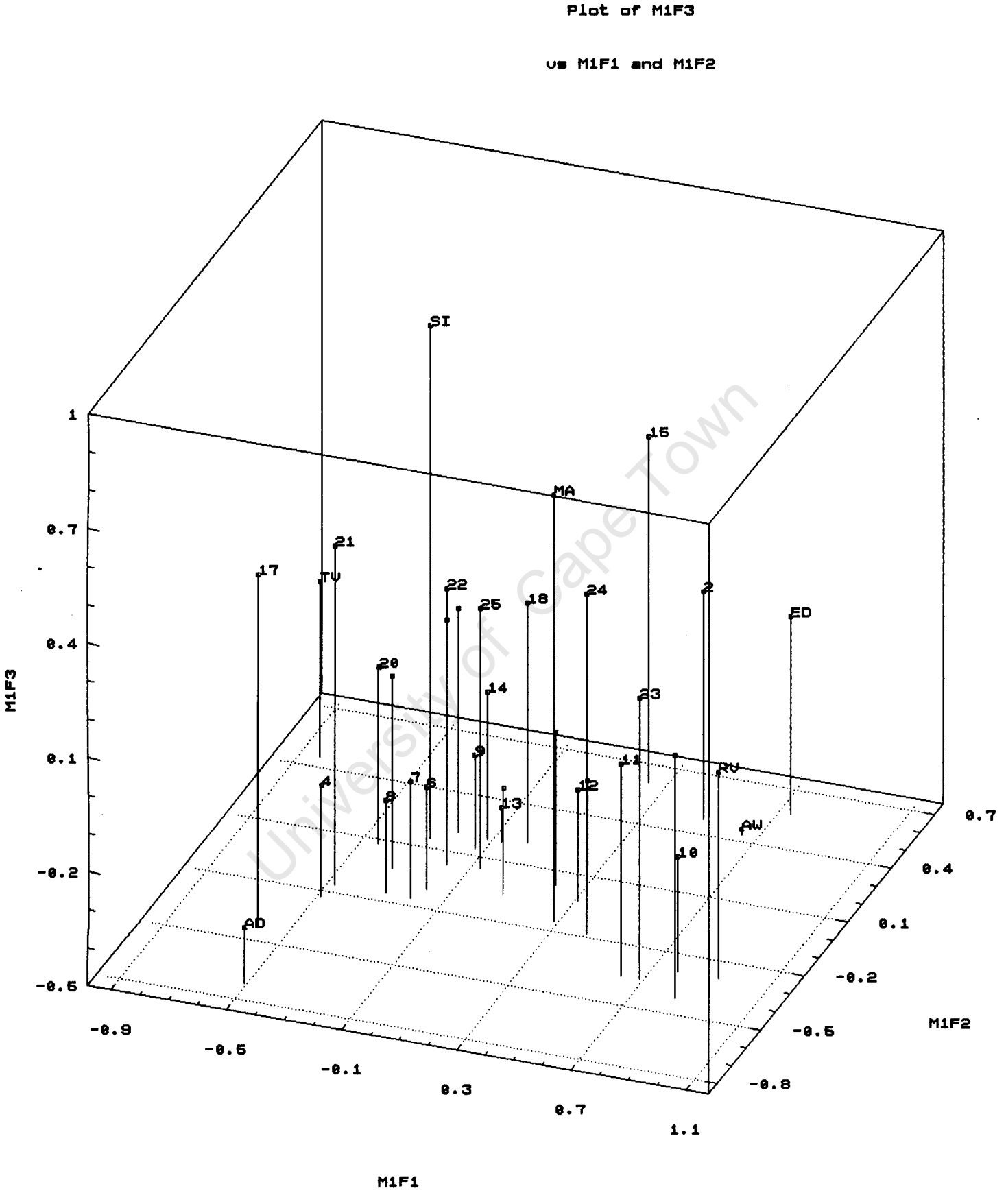
M3F1 vs M3F3 vs M3F4, M3F1 vs M3F4 vs M3F8 and even M3F2 vs M3F3 vs M3F4 substantiate the last results.

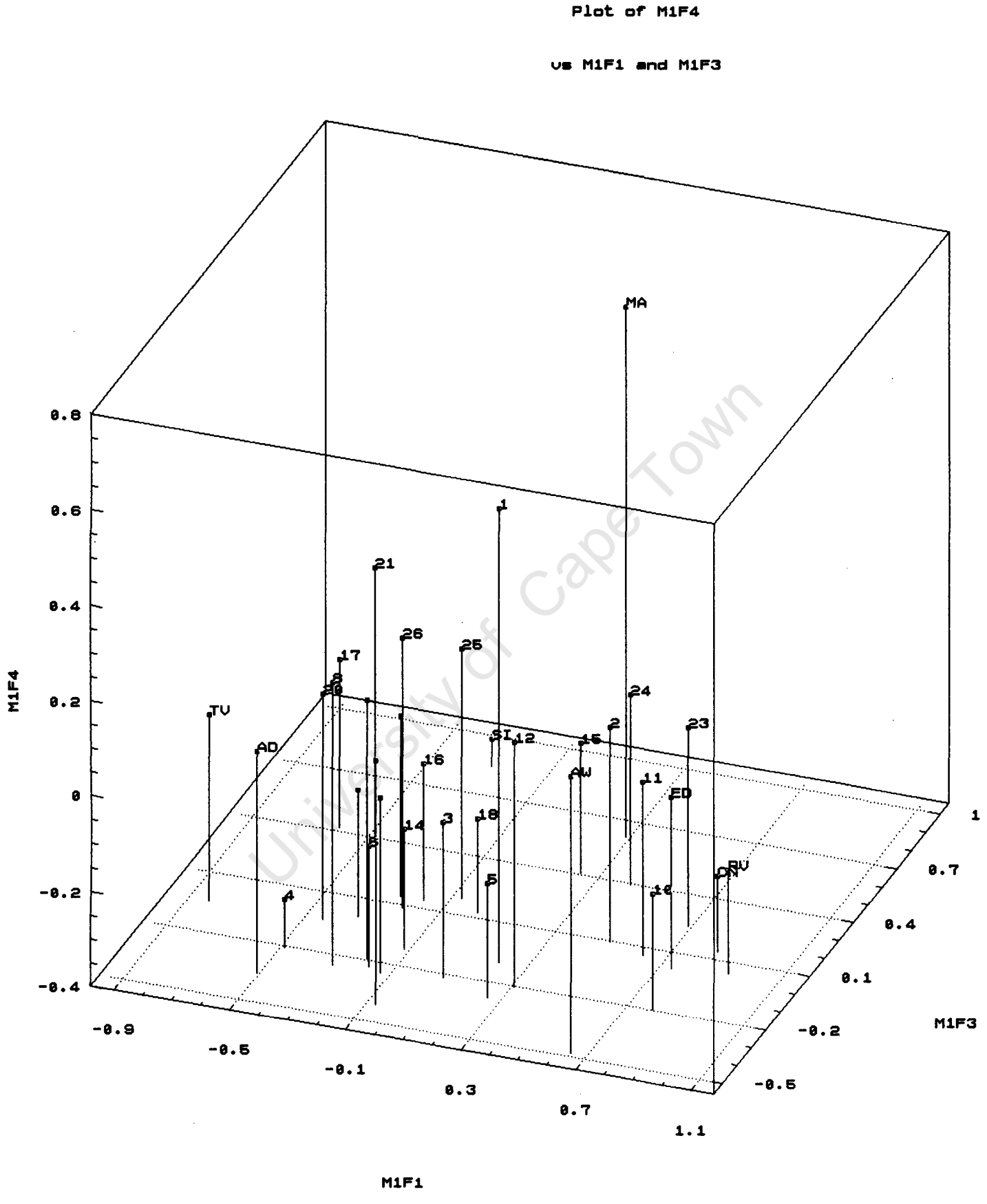


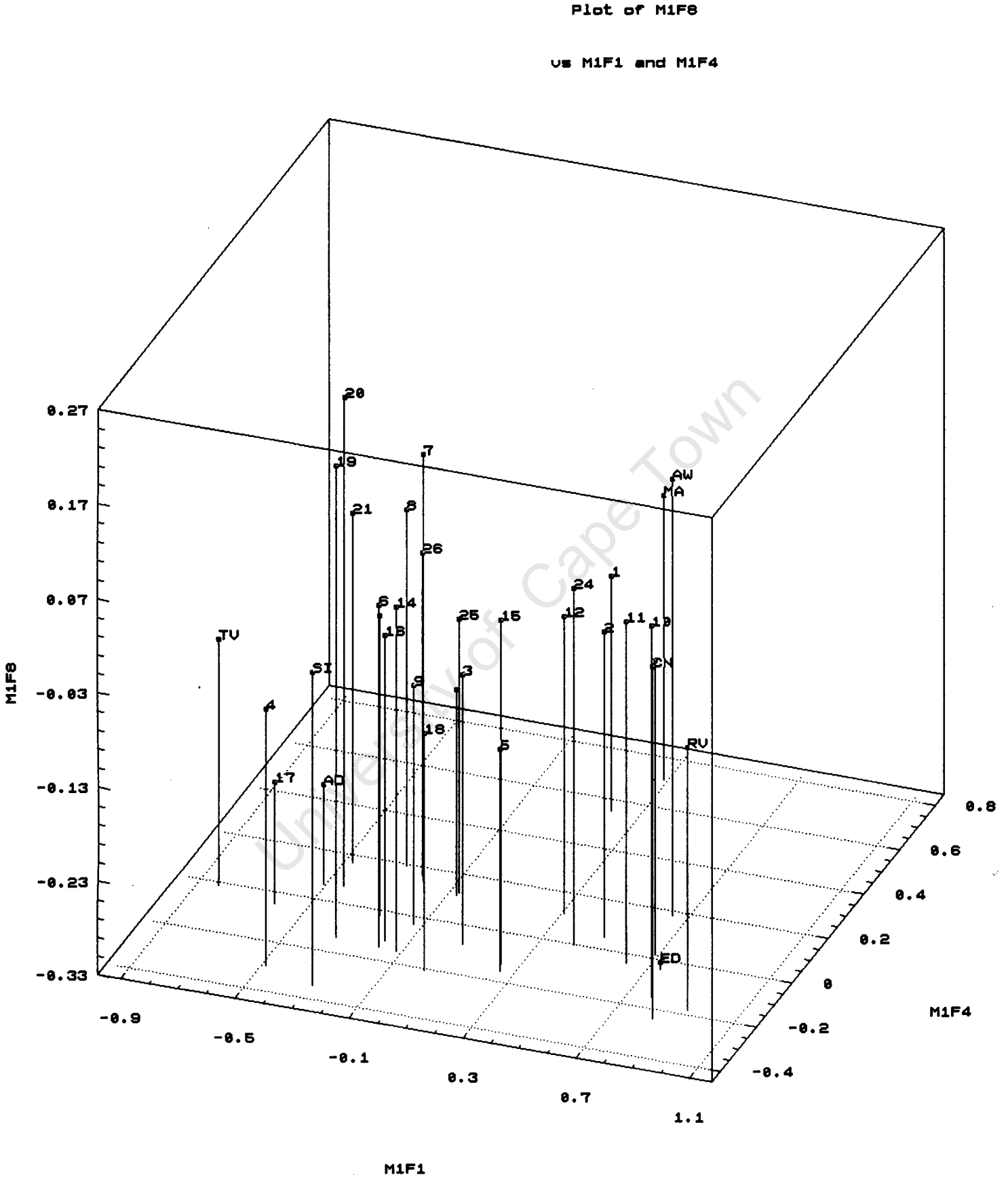


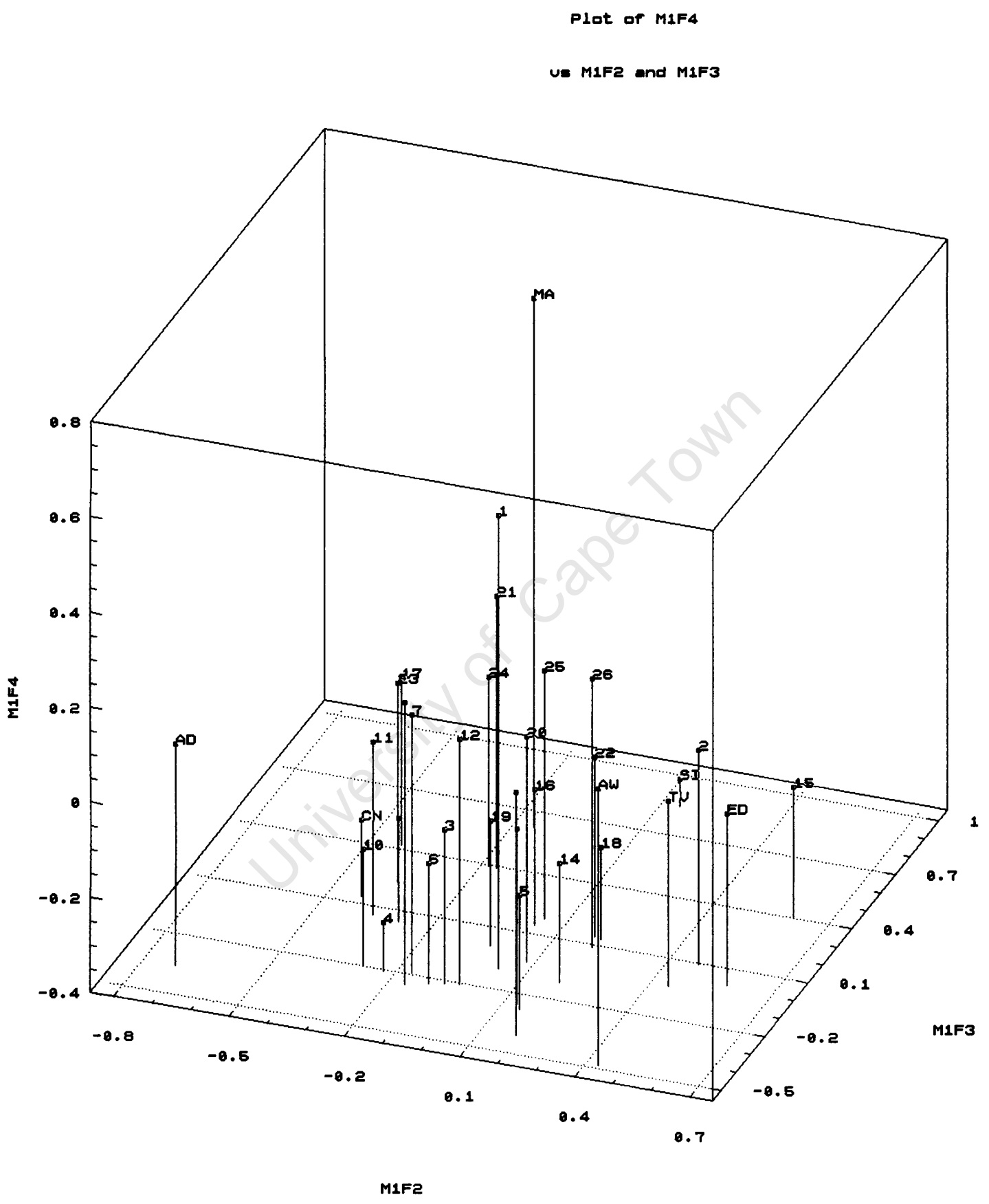


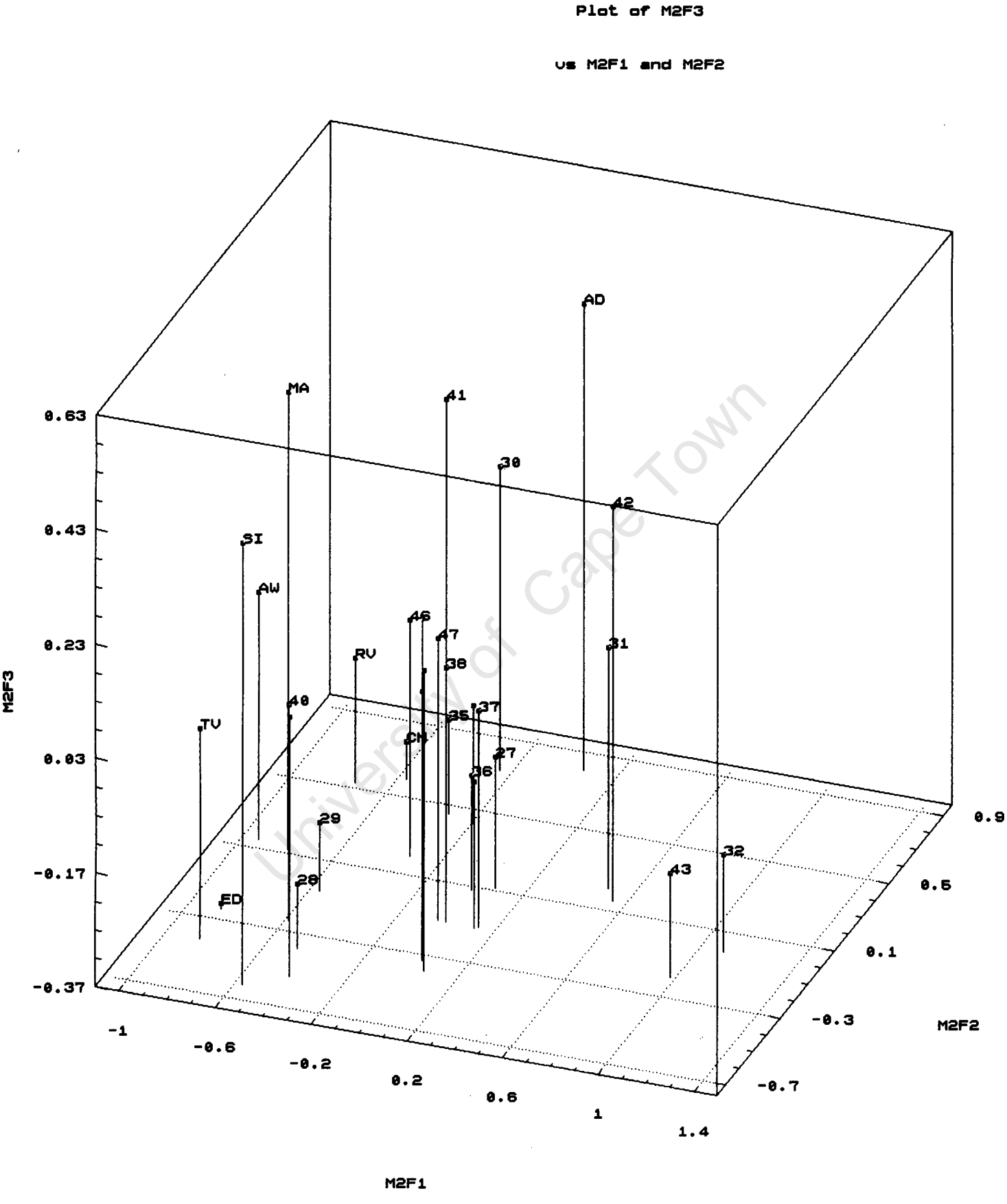


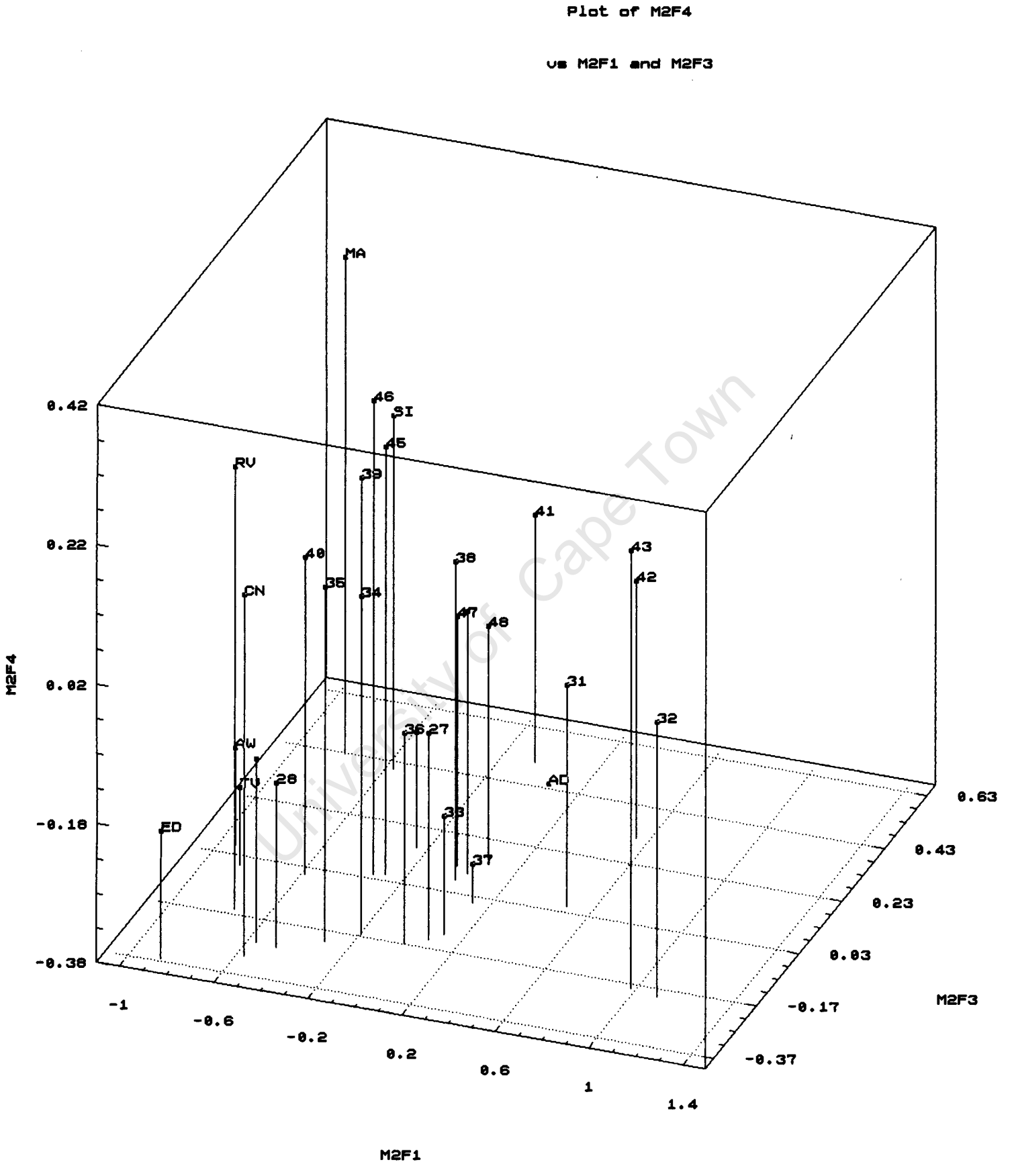

Plot of M1F3
vs M1F1 and M1F2
M1F3
M1F1
M1F2
1
0.7
0.4
0.1
-0.2
-0.5
-0.9
-0.5
-0.1
0.3
0.7
1.1
0.7
0.4
0.1
-0.2
-0.5
-0.8
SI
MA
15
21
17
TV
22
25
18
24
2
ED
20
14
23
9
11
RV
4
7
6
12
13
AW
10
AD
University of Cape Town

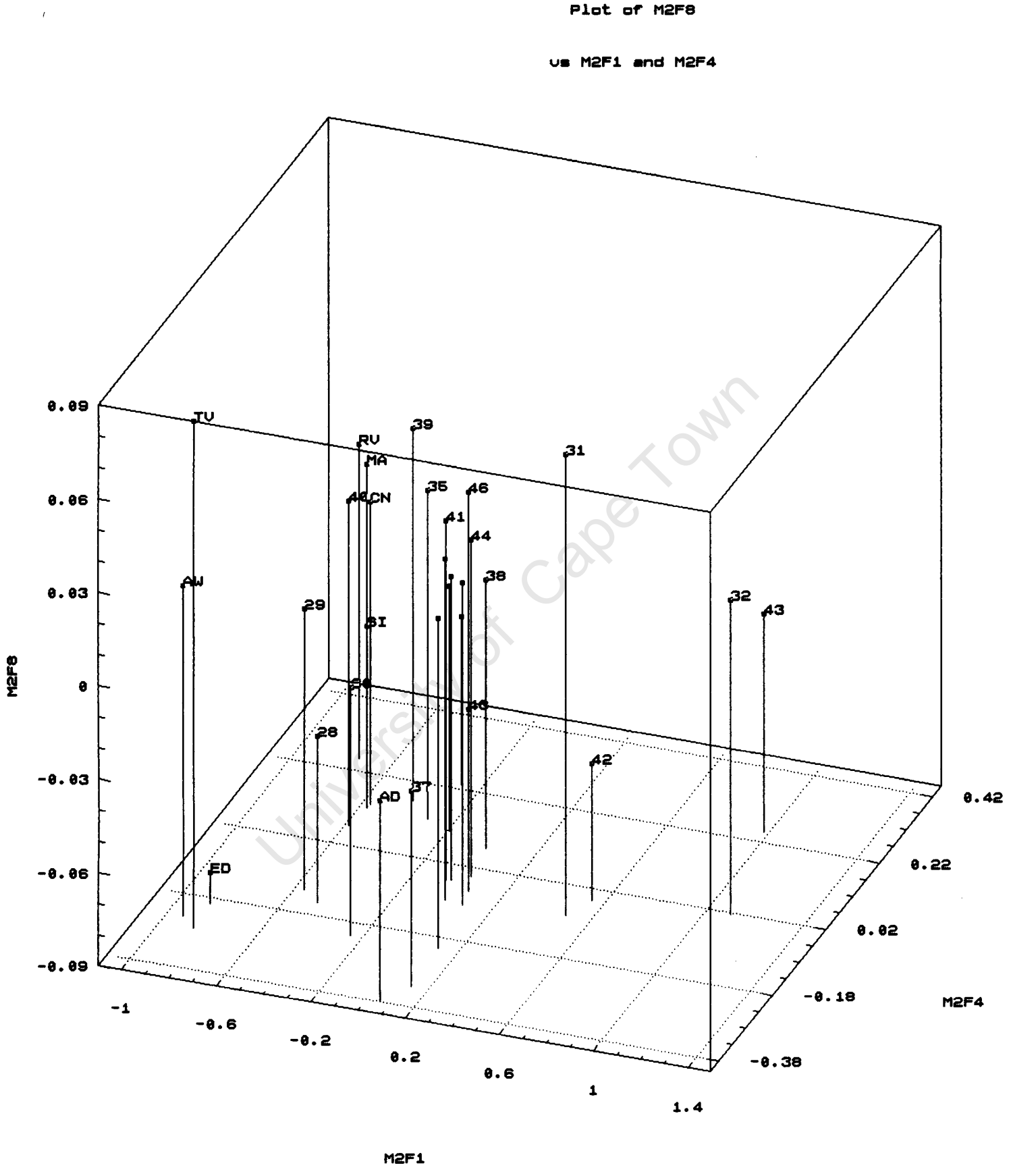


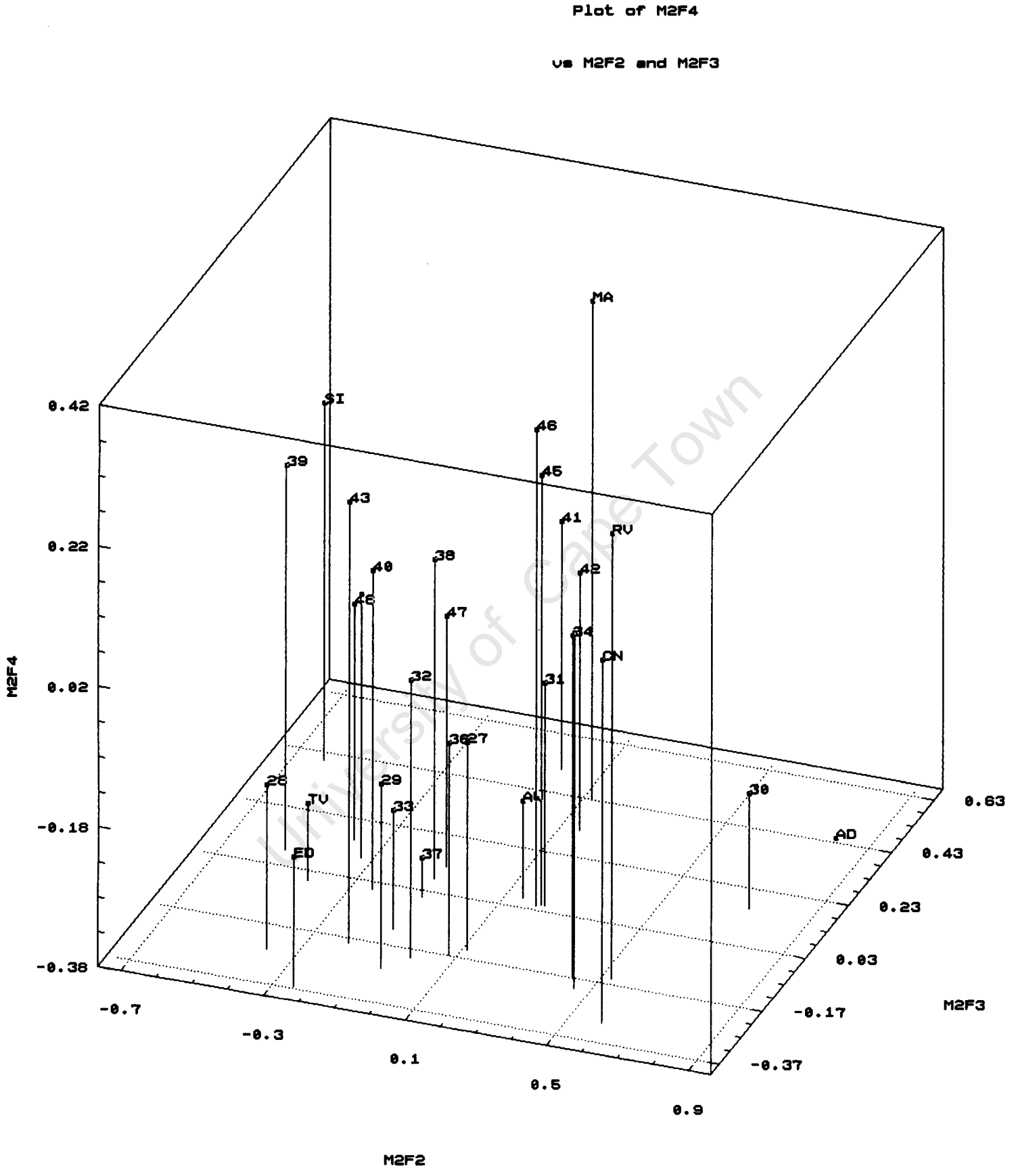

Plot of M1F8
vs M1F1 and M1F4
M1F8
0.27
0.17
0.07
-0.03
-0.13
-0.23
-0.33
M1F1
-0.9
-0.5
-0.1
0.3
0.7
1.1
M1F4
0.8
0.6
0.4
0.2
0
-0.2
-0.4
20
19
7
21
8
26
6
14
13
25
15
24
12
1
2
11
10
9
3
18
5
4
17
TV
SI
AD
AW
MA
RV
ED
University of Cape Town

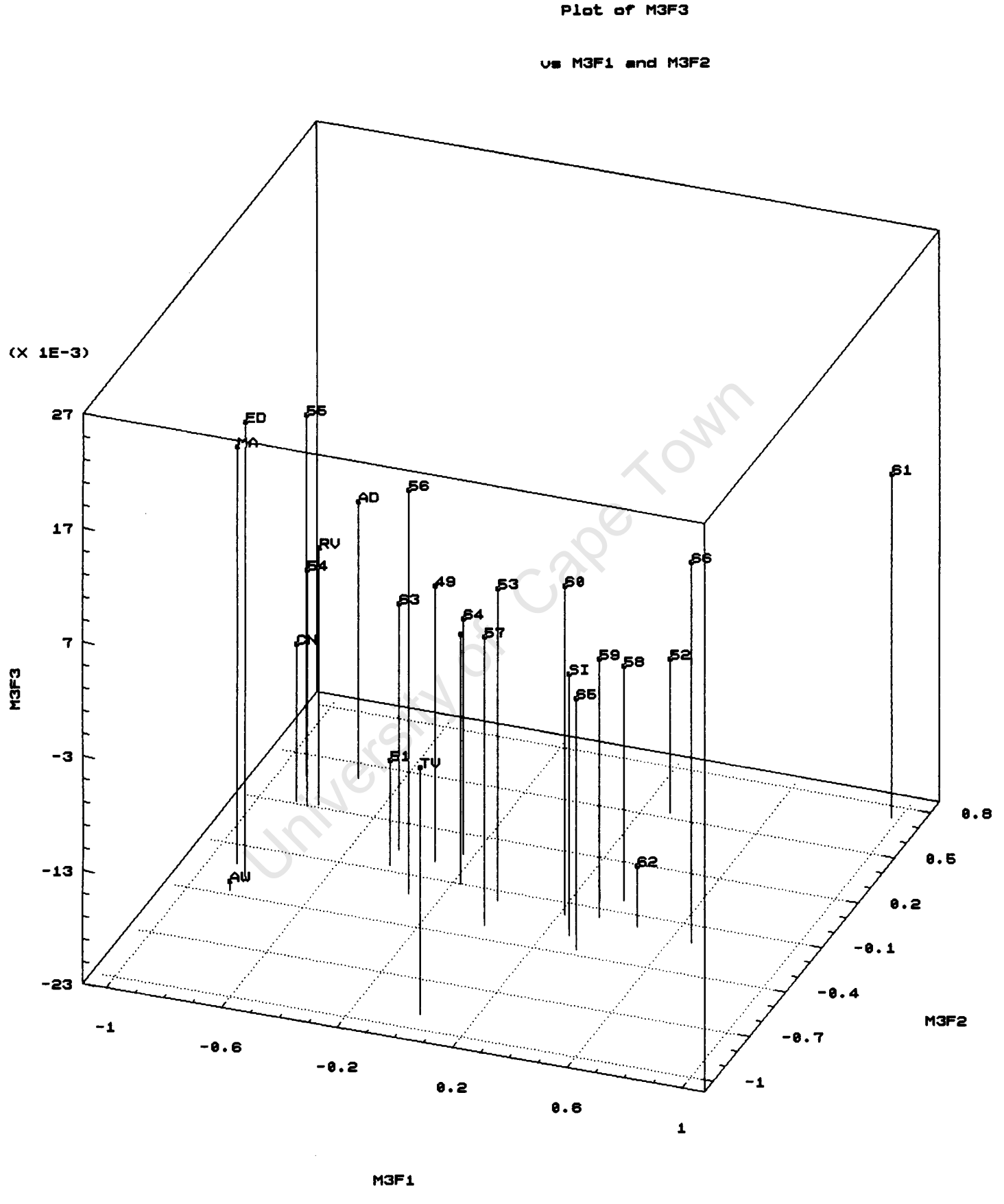


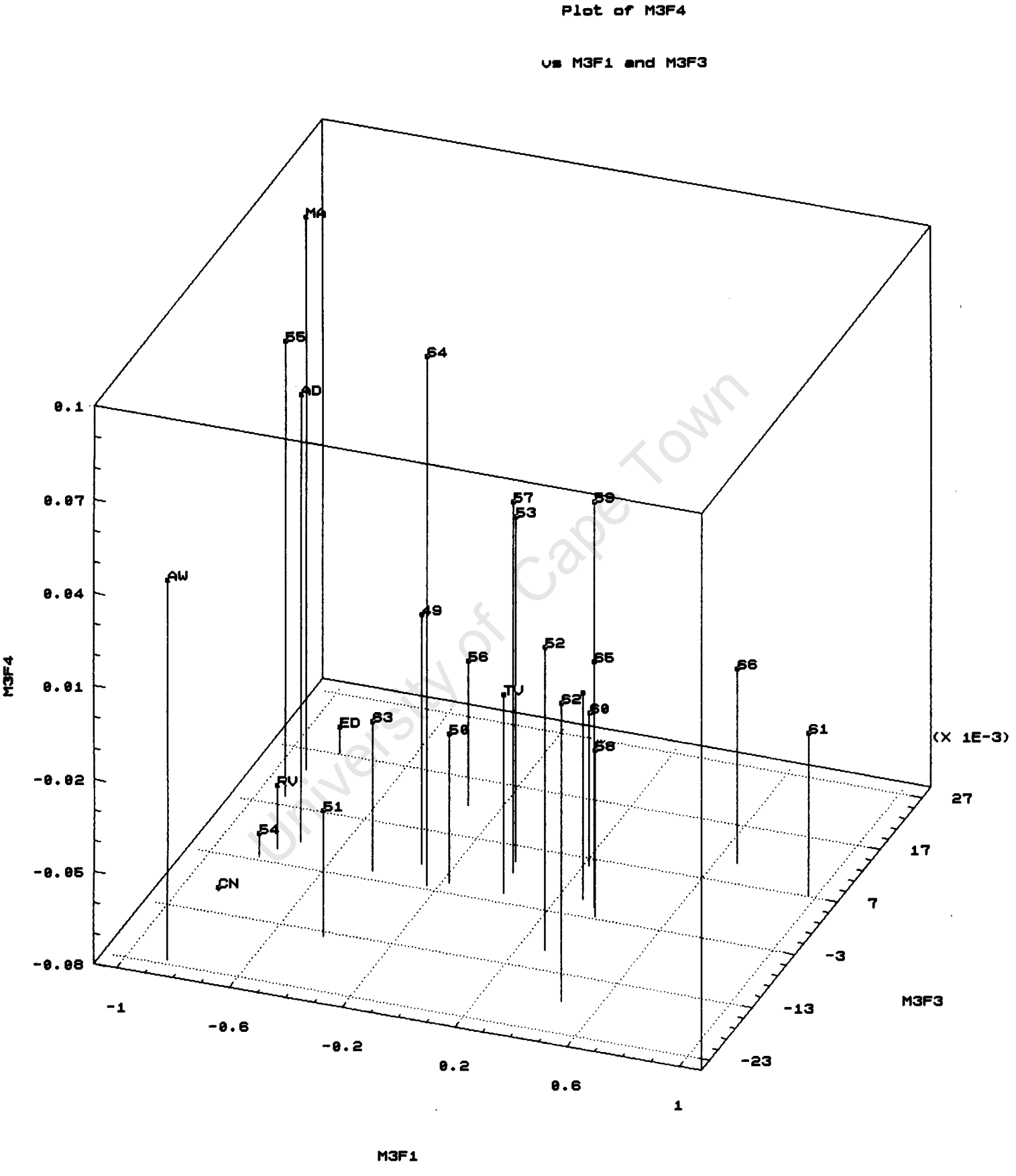


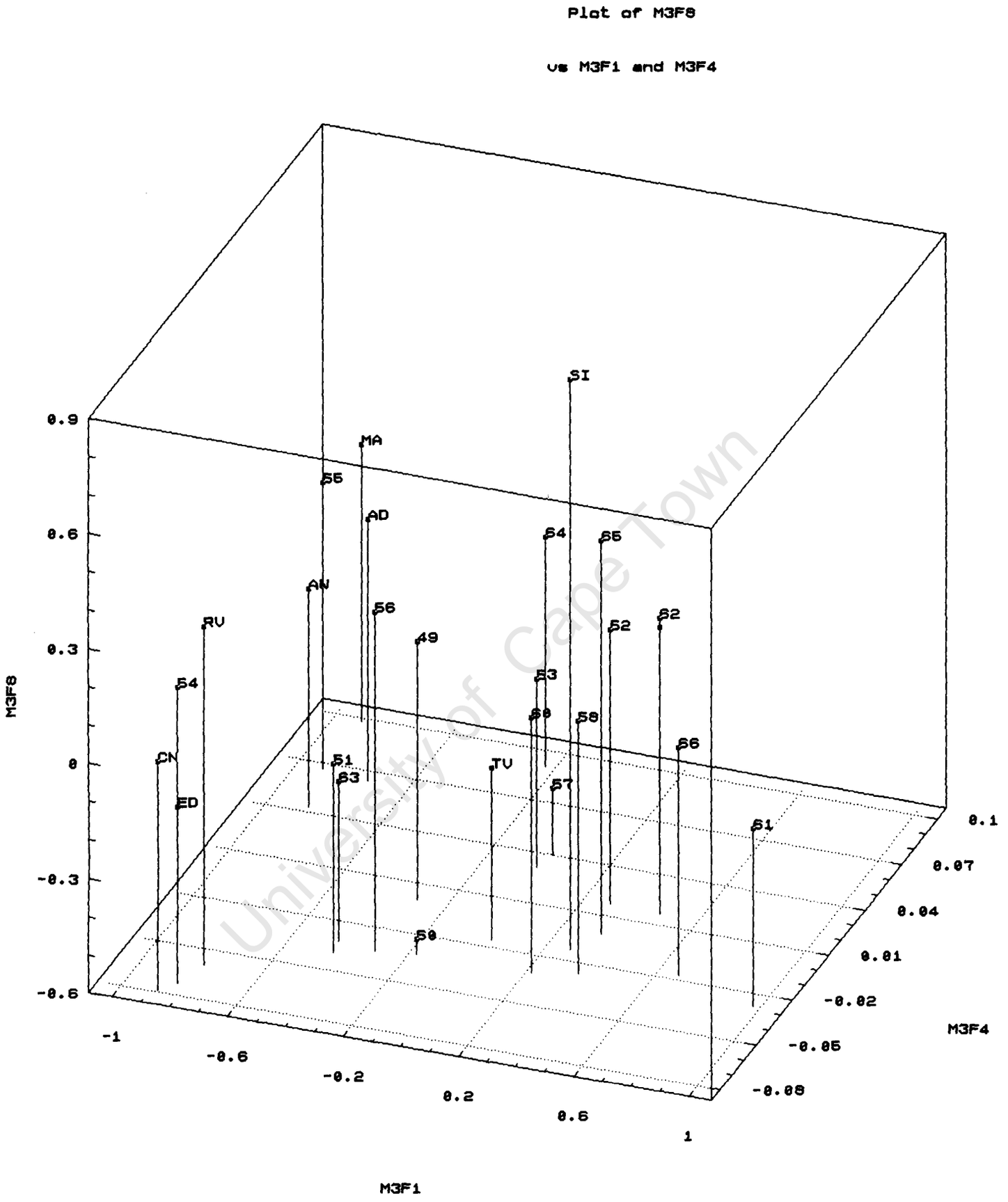


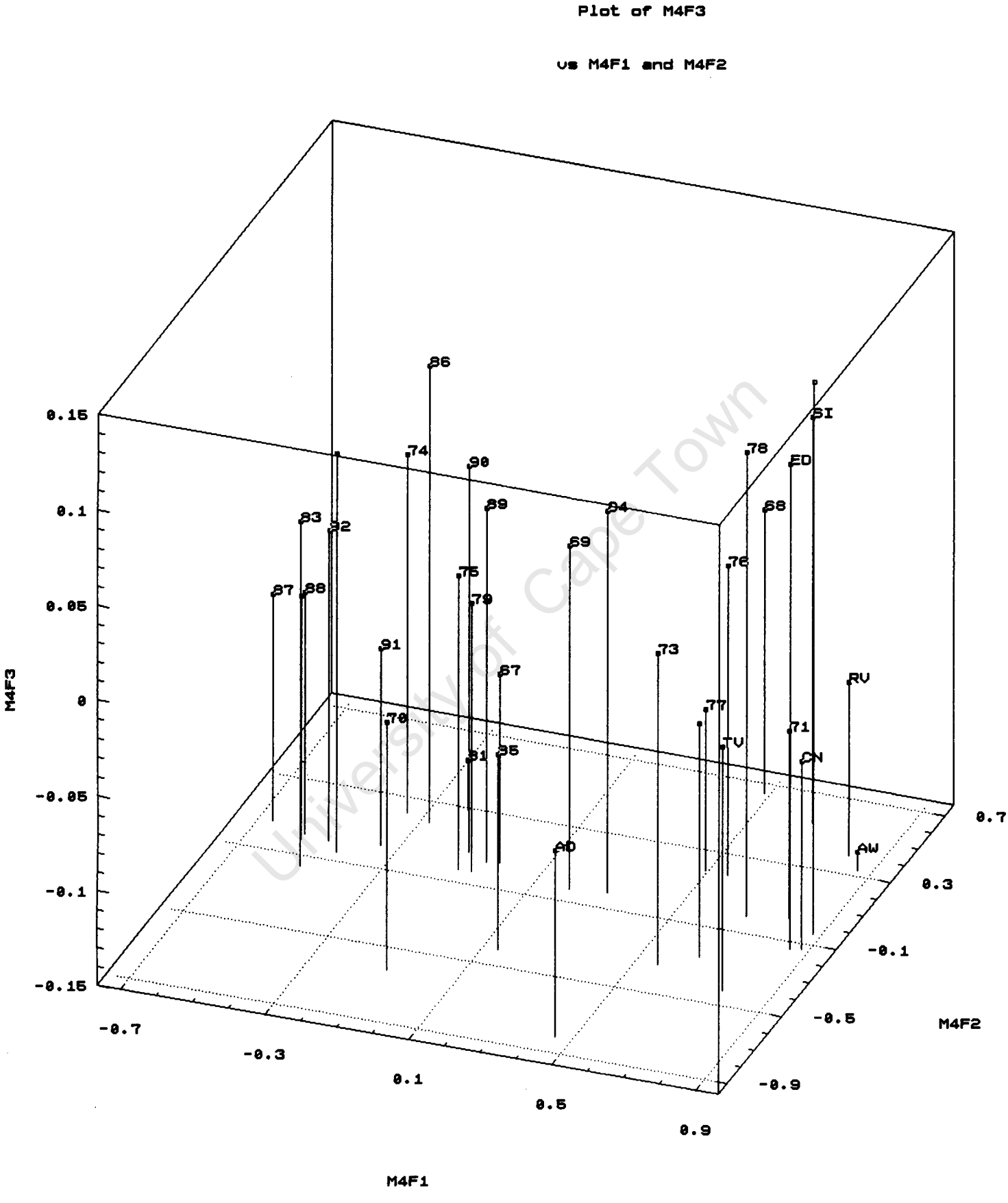


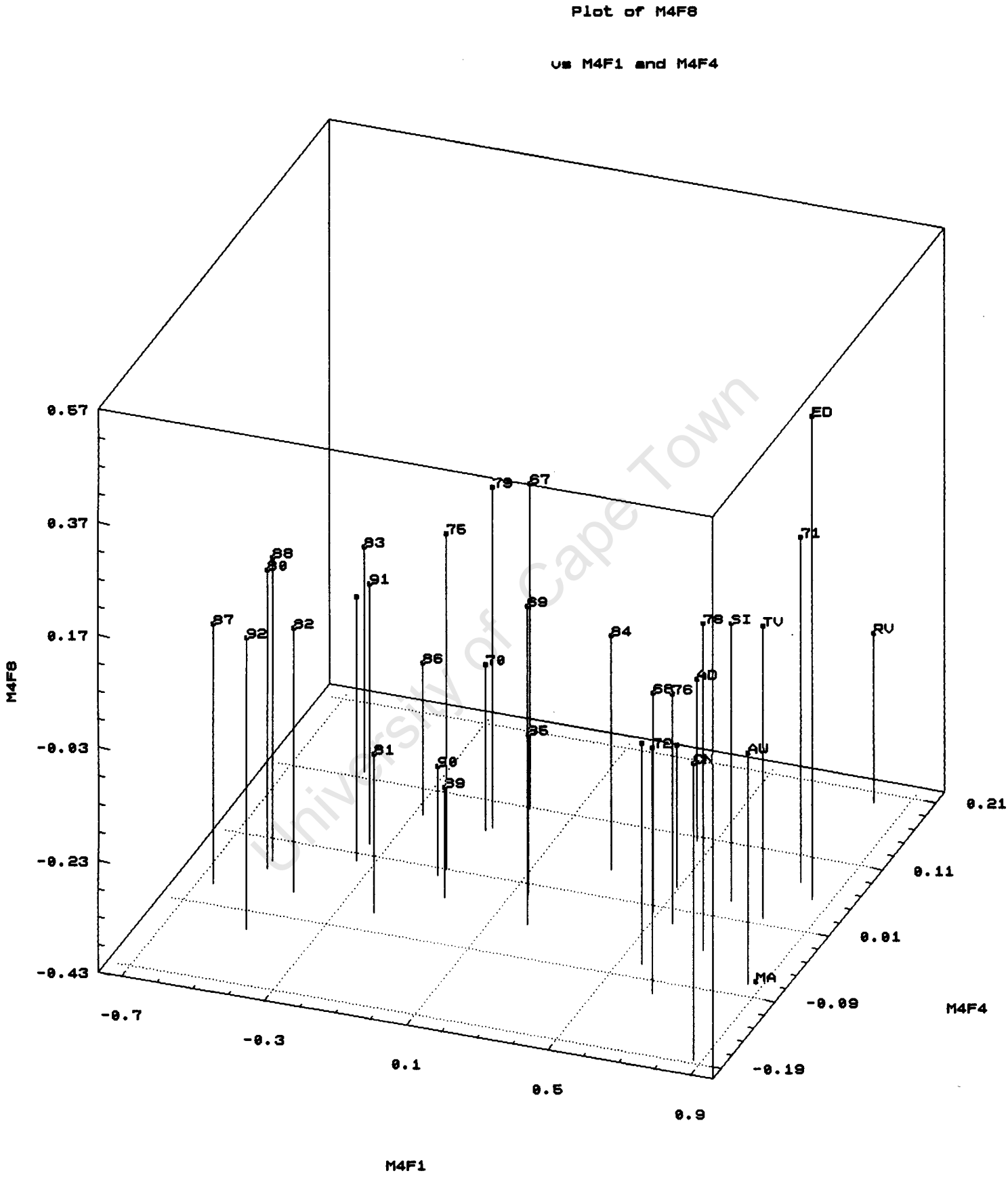


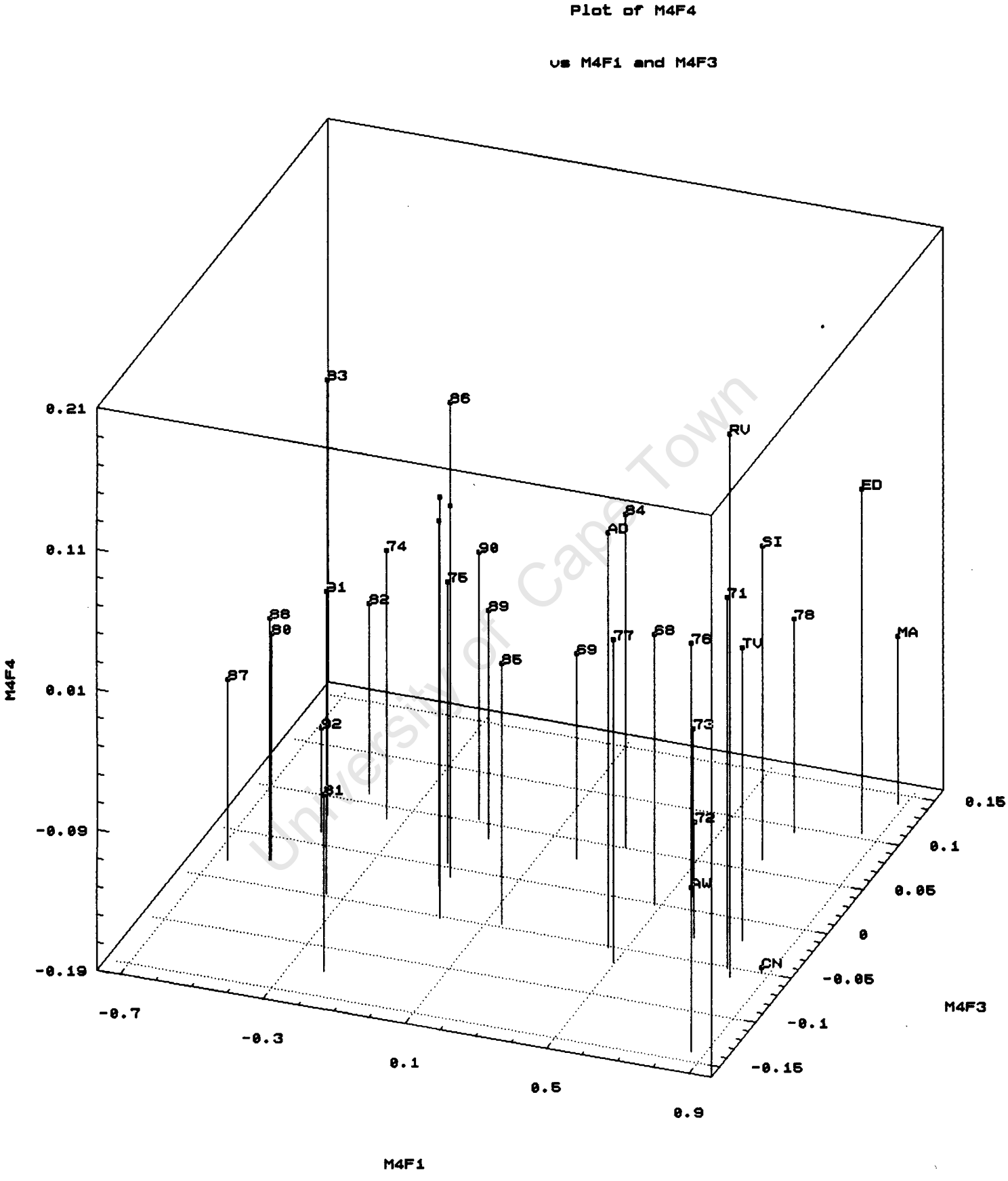

Plot of M3F3
vs M3F1 and M3F2
(X 1E-3)
27
17
7
-3
-13
-23
M3F3
-1
-0.6
-0.2
0.2
0.6
1
M3F1
0.8
0.5
0.2
-0.1
-0.4
-0.7
-1
M3F2
ED
MA
S5
AD
S6
RV
S4
S9
S3
S3
S4
S7
DN
S0
S9
S8
S2
SI
S5
S1
TV
S6
S2
AW
S1
University of Cape Town

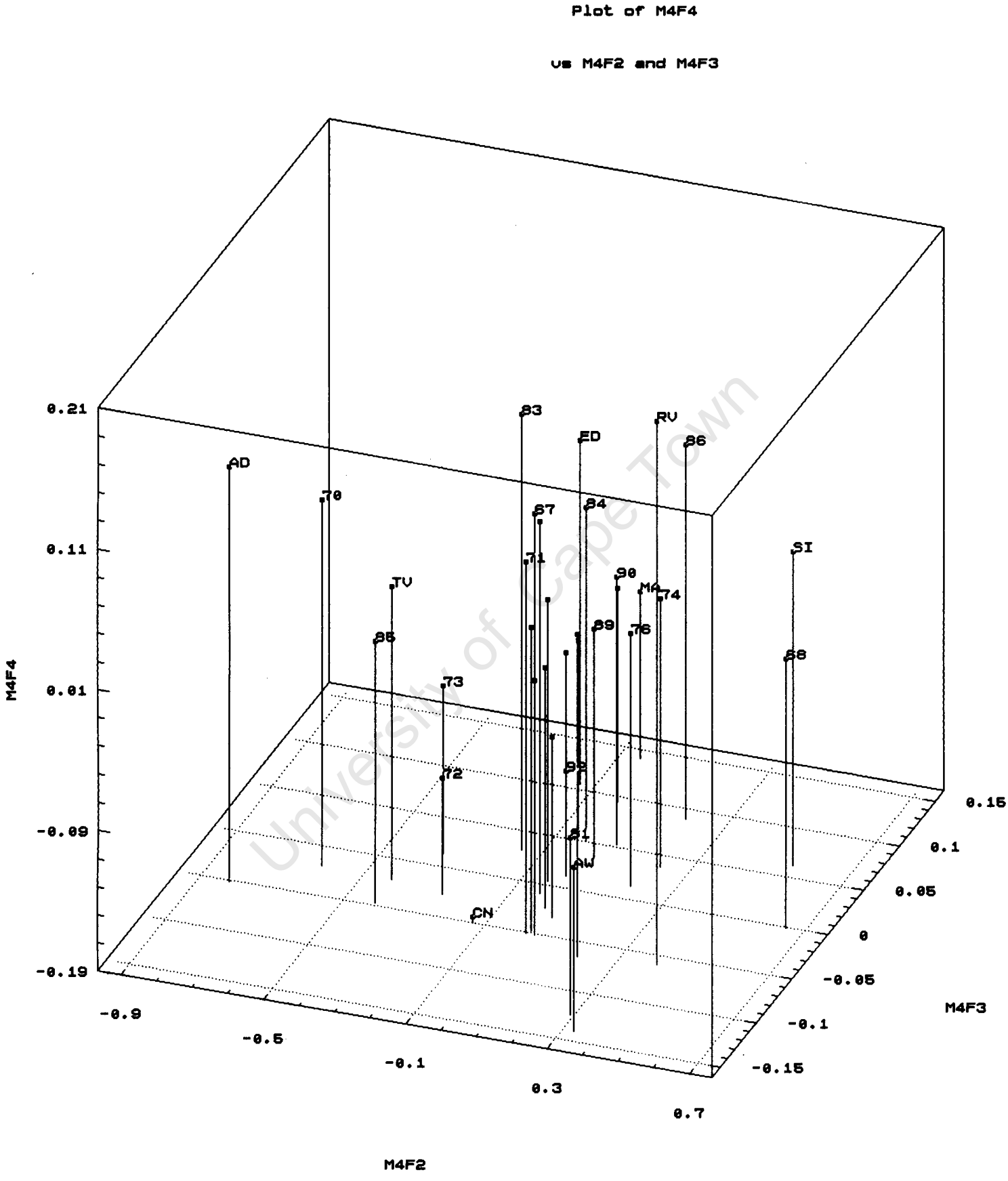



Plot of M3F8
vs M3F1 and M3F4
M3F8
M3F1
M3F4
0.9
0.6
0.3
0
-0.3
-0.6
-1
-0.6
-0.2
0.2
0.6
1
0.1
0.07
0.04
0.01
-0.02
-0.05
-0.08
SI
MA
S5
AD
S4
S5
AN
S6
S2
S2
RV
49
S4
S3
S9
S8
S6
CN
S1
S3
TV
S7
ED
S1
S0
University of Cape Town


Plot of M4F3
vs M4F1 and M4F2
M4F3
M4F1
M4F2
0.15
0.1
0.05
0
-0.05
-0.1
-0.15
-0.7
-0.3
0.1
0.5
0.9
-0.9
-0.5
-0.1
0.3
0.7
86
74
90
89
94
93
92
69
75
87
88
79
91
73
97
76
78
ED
SI
68
RV
77
71
TV
CN
70
81
85
AD
AW
University of Cape Town


Plot of M4F8
vs M4F1 and M4F4
M4F8
0.57
0.37
0.17
-0.03
-0.23
-0.43
M4F1
-0.7
-0.3
0.1
0.5
0.9
M4F4
0.21
0.11
0.01
-0.09
-0.19
ED
71
RV
TV
SI
AV
MA
78
40
76
66
72
84
69
67
79
75
83
91
88
80
87
92
82
86
70
81
90
89
35
University of Cape Town




Plot of M4F4
vs M4F2 and M4F3
M4F4
0.21
0.11
0.01
-0.09
-0.19
M4F2
-0.9
-0.5
-0.1
0.3
0.7
M4F3
0.15
0.1
0.05
0
-0.05
-0.1
-0.15
AD
70
83
ED
RV
86
87
84
71
SI
90
MA
74
TV
89
76
85
88
73
72
92
91
AW
CN
University of Cape Town

M4F1 vs M4F2 vs M4F3: Observations 83, 87 and 88 stand out in this plot as well as in M4F1 vs M4F4 vs M4F8 and M4F1 vs M4F3 vs M4F4 (see observation 83) but they vanish in M4F2 vs M4F3 vs M4F4.

The graphical representations thus substantiate the numerical analysis.

The question is: What will happen if these outliers/influential observations are removed and the LDF is repeated ? Removing observations 10, 43, 61 and 87 arbitrarily (those observations of highest relative variance contribution in each group or deviant position with respect to group in the graphical display) resulted in an apparent classification rate of correct/incorrent: 96.16%/3.4% which represents only a small change in comparison to previous results.

One could expect that removing more than only these 4 observations should have a rather more drastic effect eventually.



CLASSIFICATION - MEDIA DATA (10, 43, 61, 87 deleted)

| | | Predicted | | | |
|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 |
| | 1 | 25 | 0 | 0 | 0 |
| Actual | 2 | 0 | 18 | 2 | 1 |
| | 3 | 0 | 0 | 17 | 0 |
| | 4 | 0 | 0 | 0 | 25 |

At this stage on can accept that although the SVD analysis together with the 3-dimensional biplots will not necessarily indicate the misclassified

observations which was not the idea in the first place, but it can give an indication of these observations which may have relatively large influence on the analysis or are just ordinary outliers. Again it must be emphasized that influential observations may be either good - i.e. stabilizing a model, or bad, i.e. relating badly to the consumer by being not efficient in the predictive sense of the word.

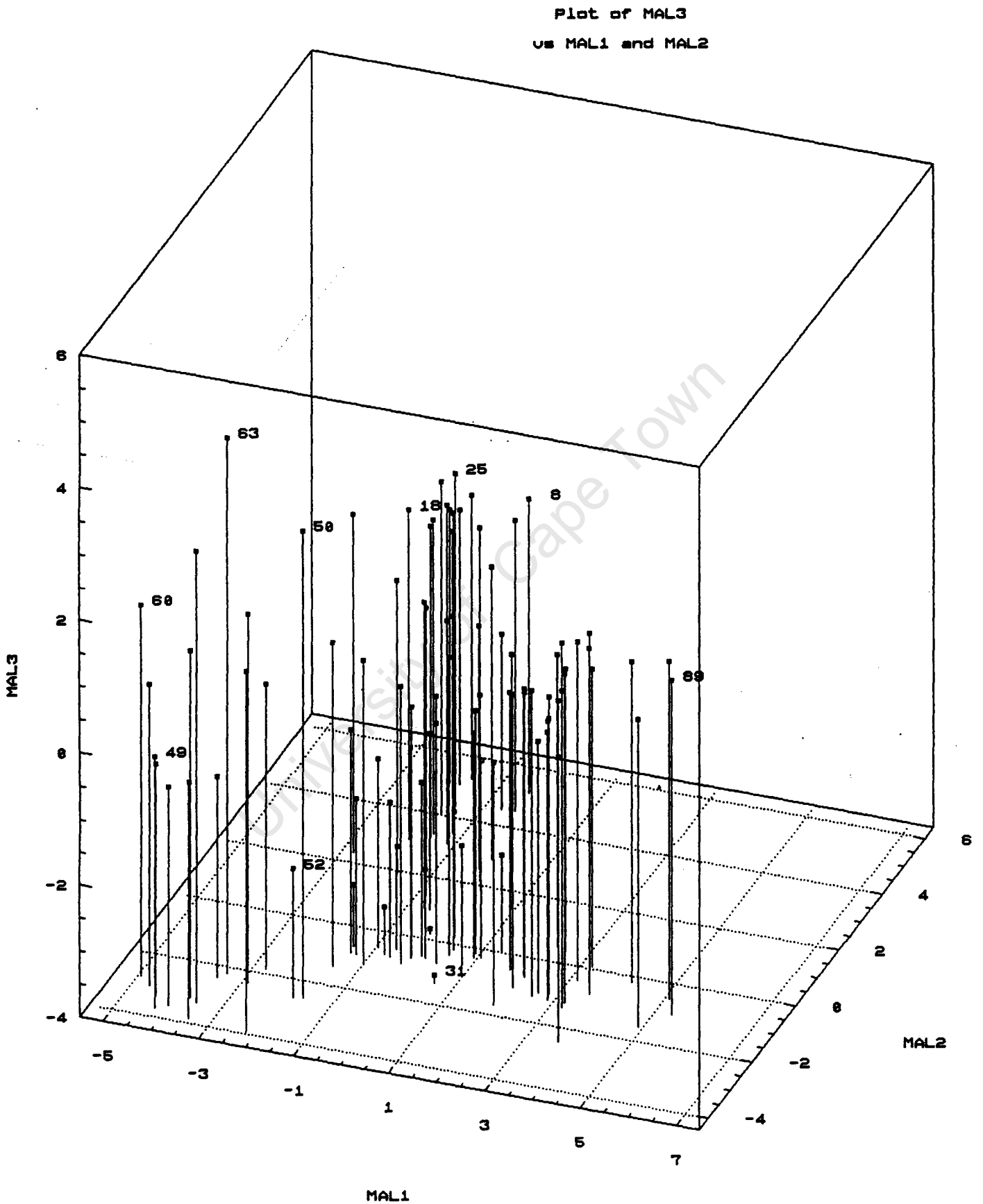
## 5.7 COMPARISONS TO OTHER TECHNIQUES AND EXTENSIONS

### 5.7.1 Introduction

In this section we discuss the problem of unequal covariance matrices where the assumption of equal covariance matrices are basic to the application of the LDF approach.

In the first place how does one know whether these matrices are different ? If so, what can one do about it ?

With respect to the last question posed: One thing I did try was to apply the bounded-influence weighting of Mallows on the independent variables. Then I used these bounded-influence weighted variables to estimate a bounded-influence covariance matrix. This was done on each class group and then I found the pooled covariance matrix as usual - now a bounded-influence pooled covariance matrix. Then I carried on as usual with the LDF. On the next 2 pages the 3-dimensional media biplots can be seen. The one graph has a $\tau = 0,15$ smooth and the other a $\tau = 0,05$ smooth. As one can see the effect was that the difference between groups are smoothed out considerably so that one can not distinguish between



groups as easily as before. Note that the example refer to the Media data of the last section where we know the classification was quite good. Refer also to the original 3-dimensional biplots with respect to these data.

If one thinks about it - smoothing out differences is just the opposite to what one really tries to do in a discriminant analysis. That is why we can see that the 5% plot (MA5*) gives a better picture than the 15% plot (MAL*).

### 5.7.2 Test for equal covariance matrices

To test

$$H_0: \Sigma_1 = \Sigma_2 = \ldots = \Sigma_k$$

against a suitable alternative we use the unbiased estimates $S_i$ of $\Sigma_i$ based on $n_i$ degrees of freedom where $n_i = N_i - 1$ (Morrison, 1976). When these $S_i$'s are pooled we find the usual

$$S = \frac{1}{n_i} \Sigma n_i S_i$$

If we define the test statistic

$$M = \Sigma n_i \ell n|S| - \Sigma n_i S_i$$

then Box (1949) has shown that if for fairly large $n_i$ the scale factor

$$c^{-1} = 1 - \frac{2p^2 + 3p - 1}{6(p+1)(k-1)} \left( \sum_{i=1}^{k} \frac{1}{n_i} - \frac{1}{\Sigma n_i} \right)$$

is introduced then $Mc^{-1}$ is approximately distributed as $\chi^2$ with $\frac{1}{2}(k-1)p(p+1)$ degrees of freedom.

If $k = 2$ the optimal test for

$$H_0: \Sigma_1 = \Sigma_2 = \ldots \Sigma_k$$

is Pillai's V where

$$V = \text{trace}(S_1 + S_2)^{-1} S_2$$

Appropriate references are John (1971) and Pillai (1957).

If we apply the tests for equal/unequal covariance matrices on the examples of the previous section we find the following:

| | Box | Pillai trace |
|---|---|---|
| Bank | $\chi^2 = 64.5$; significant | 3.861 significant* |
| | (d.f. = $\frac{1}{2}(4)(4+1)(2-1) = 10$) | |
| Media | $\chi^2 = 680$; significant | - |
| | (d.f. = $\frac{1}{2}(8)(8+1)(4-1) = 180$) | |

*($n_1 = 21$; $n_2 = 25$; $m = \frac{1}{2}(n-p-2) = \frac{1}{2}(21-4-2) = 7.5 \doteq 8$

$n = \frac{1}{2}(m-p-2) = \frac{1}{2}(25-4-2) = 9.5 \doteq 10$

so that $n = V_{0.05} = 2{,}31$ and $V_{0.01} = 2.44$)

From these results it is clear that the straight forward application of the LDF on the given examples was not correct in the first place in case any inferences had to be drawn from the results. The good classification rate that was obtained is then an indication of the strong differences between these different classes.

It is interesting at this point to determine the Box test statistic on the well-known Iris data. We find that $\chi^2 = 13.1106$ with df = $\frac{1}{2}p(p+1)(k-1) = \frac{1}{2}(4)(4+1)(3-1) = 20$ which shows that the application of the LDF was well founded in that case.

Accordingly we are going to compare our results of the last section to the results of some other techniques, but first some information about some of these techniques.

### 5.7.3 Ridge regularization and flexible discriminant analysis

Friedman (1989) described the regularization in discriminant analysis in order to indicate how the individual class sample covariance matrices can be replaced by their average i.e. let

$$\hat{\pi}_k = \frac{V_k}{V}$$

where $V_k$ is equal to $n_i$ and $V = n = \sum_{i}^{k} n_i$ if equal weights are attached to all observations. Then if $S_k$ here represents the sum of squares of errors

$$\hat{\Sigma} = S/V = \Sigma S_k/V$$

The choice between linear and quadratic discriminant analysis can be less restricted by introducing the regularization parameter $\lambda$ such that

$$\hat{\Sigma}_k(\lambda) = S_k(\lambda)/W_k(\lambda)$$

with

$$S_k(\lambda) = (1 - \lambda)S_k + \lambda S$$

The regularization parameter $\lambda$ can now take on any values such that $0 \leq \lambda \leq 1$, so that the shrinkages of the individual covariance matrices can be controlled. If $\lambda = 0$ we have the QDA and if $\lambda = 1$ we have the LDA. If the class covariances differ substantially a better performance can be obtained by reducing $\lambda$ away from 1. Friedman goes further and mention some more regularization steps.

Troskie and Conradie (1986) derived the exact distribution of the condition number $\lambda_1/\lambda_p$ of an estimated covariance matrix $\Sigma$ of a multivariate $n(\mu, \Sigma)$ distribution. The expression was very complicated.

Under the assumption that $\Sigma = \sigma^2 I$ they showed that critical values for $\lambda_1/\lambda_p$ of the order of 20 appear to be large. If $\lambda_1 > \lambda_2 \ldots > \lambda_p$ are the population characteristic roots of $\Sigma$ then they show that the density of $\lambda_1/\lambda_p$ is $\frac{\lambda_1}{\lambda_p} F((\frac{1}{2}(n-p+1), \frac{1}{2}(n-p+1))$.

Assuming therefore that $\lambda_1/\lambda_p = 20$ we see that a cutoff value of

$$20F^{\alpha}(\tfrac{1}{2}(n-p+1),\ \tfrac{1}{2}(n-p+1))$$

seems reasonable. If $n \doteq 100$ and $p = 7$, $F^{0.01} \approx 2$ so that a cutoff value is approximately 40. We have decided to use a cutoff value of 30 when using a ridge correction for the condition number in the following way. Compute the characteristic roots of the correlation matrix and let $\lambda_1/\lambda_p$ be the condition number of the correlation matrix where the latter is more stable than the covariance matrix. Since the ridge constant k is $0 < k < 1$ the largest root will really not be effected by adding a small constant k.

Thus we suggest to use a ridge correction if

$$\frac{\lambda_1}{\lambda_p + k} \geq 30$$

or

$$k \doteq \frac{\lambda_1}{30} - \lambda_p$$

Flexible discriminant analysis is described by Trevor Hastie, Tibshirani and Buja (1992) as a tool for richer non-linear classification. Other related references are Gnanadesikan and Kettering (1989), Friedman (1987, 1991) and Hastie and Tibshirani (1990). Hastie et al summarized Friedman's 1991 procedure.

Suppose we have response variables $Y_i$, $i = 1,\ldots,k$ and a vector of predictors $\underline{X}$, then

$$\eta_i(\underline{X}) = \beta_{0i} + \sum_{m=1}^{M} \beta_{mi} h_m(\underline{X}),\ i = 1,\ldots,k$$

The coefficients $\beta_{ij}$ may be determined by means of least squares or penalized least squares or a $L_p$-norm procedure - whatever the choice may be. The basic functions $h_m(\underline{X})$ are chosen adaptively.

The MARS (Multivariate adaptive regression splines) proposal of Friedman (1991) then is a procedure for adaptive non-parametric regression. The regression function is approximated by

$$f(x) = \beta_0 + \sum_{m=1}^{M} \beta_m \prod_{k=1}^{Km} h_{km}\left(x_{v(k,m)}\right)$$

where $x_i$, $i = 1,\ldots,p$ are predictors and $v(k,m)$ is the index of the predictor used in the kth term of the mth product. The basic functions $h_{km}$ are defined in pairs

$$h_{km}(x) = [x - t_{km}]_+$$
$$h_{k,m+1}(x) = [t_{km} - x]_+$$

for m an odd integer, where the knot value $t_{km}$ is one of the unique values of $x_{v(k,m)}$. The typical model term is now denoted by $H_j(x)$ and the model is built up in a forward stepwise manner with backward pruning steps as necessary when some maximium model size is reached. The best fitting model measured by the generalized cross validation criterion is chosen. The degree is a parameter of the procedure. First degree models are additive, whereas 2nd degree models allow pairwise interactions. Full details of Mars may be found in Friedman (1991).

Hastie et al generalized MARS for incorporation into the adaptive discriminant procedure - coding the procedure from scratch and allowing for multiple response variables. They defined each of the coefficients $\beta_m$ as a k vector together with k response variables. The residual sum of squares and generalized cross validation criterion however, now involve sums over k response variables.

Given these I compared the results for both the Bank and Media data with respect to apparent classification error rate (AER).

BANK DATA

Linear discriminant analysis:

| | | predicted classification | | |
|---|---|---|---|---|
| | | 1 | 2 | |
| classification | 1 | 18 | 2 | |
| | 2 | 3 | 23 | AER = 0.1087 |

Quadratic discriminant analysis:

| | | predicted classification | | |
|---|---|---|---|---|
| | | 1 | 2 | |
| classification | 1 | 19 | 2 | |
| | 2 | 2 | 23 | AER = 0.0870 |

Quadratic discrimination analysis with penalty ridge on each covariance matrix:

| | | predicted classification | | |
|---|---|---|---|---|
| | | 1 | 2 | |
| classification | 1 | 20 | 12 | |
| | 2 | 1 | 13 | AER = 0.2826 |

Quadratic discrimination analysis - A robust estimate using 3000 samples of size 5 each:

| | | predicted classification 1 | 2 | |
|---|---|---|---|---|
| classification | 1 | 19 | 2 | |
| | 2 | 2 | 23 | AER = 0.0870 |

Flexible discriminant analysis - method ≡ polyreg; degree = 1:

| | | predicted classification 1 | 2 | |
|---|---|---|---|---|
| classification | 1 | 18 | 1 | |
| | 2 | 3 | 24 | AER = 0.0870 |

Flexible discriminant analysis - method ≡ polyreg; degree = 2:

| | | predicted classification 1 | 2 | |
|---|---|---|---|---|
| classification | 1 | 19 | 1 | |
| | 2 | 2 | 24 | AER = 0.0652 |

## ADVERTISING MEDIA DATA (Covariance matrices unequal)

Linear discriminant analysis (not really valid)

| | | predicted classification | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | |
| classification | 1 | 26 | 0 | 0 | 0 | |
| | 2 | 0 | 19 | 0 | 0 | |
| | 3 | 0 | 2 | 18 | 0 | |
| | 4 | 0 | 1 | 0 | 26 | AER = 0.0326 |

Quadratic discriminant analysis:

| | | predicted classification | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | |
| classification | 1 | 26 | 0 | 0 | 0 | |
| | 2 | 0 | 22 | 0 | 0 | |
| | 3 | 0 | 0 | 18 | 0 | |
| | 4 | 0 | 0 | 0 | 26 | AER = 0.0 |

Quadratic discriminant analysis with penalty ridge on each covariance matrix:

| | | predicted classification | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | |
| classification | 1 | 26 | 0 | 0 | 0 | |
| | 2 | 0 | 21 | 0 | 0 | |
| | 3 | 0 | 1 | 18 | 0 | |
| | 4 | 0 | 0 | 0 | 26 | AER = 0.0109 |

Quadratic discriminant analysis - A robust estimate using 3000 samples of 9 each:

| | | predicted classification | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | |
| classification | 1 | 26 | 0 | 0 | 0 | |
| | 2 | 0 | 21 | 0 | 0 | |
| | 3 | 0 | 1 | 18 | 1 | |
| | 4 | 0 | 0 | 0 | 25 | AER = 0.0217 |

Flexible discriminant analysis - method ≡ polyreg; degree = 1:

| | | predicted classification | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | |
| classification | 1 | 24 | 0 | 0 | 0 | |
| | 2 | 0 | 20 | 0 | 0 | |
| | 3 | 0 | 1 | 18 | 0 | |
| | 4 | 0 | 1 | 0 | 26 | AER = 0.0217 |

Flexible discriminant analysis - method ≡ polyreg; degree = 2:

| | | predicted classification | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | |
| classification | 1 | 26 | 0 | 0 | 0 | |
| | 2 | 0 | 22 | 0 | 0 | |
| | 3 | 0 | 0 | 18 | 0 | |
| | 4 | 0 | 0 | 0 | 26 | AER = 0.0 |

Although this is not a simulation study it is interesting to note the performance of the ordinary quadratic discriminant technique and then in particular that of

the flexible discriminant analysis - degree 2.

## 5.8 SUMMARY AND CONCLUSIONS

If this chapter a brief review was given of measures to determine the degree of collinearity inherent in a data matrix. A further review was given of biasing in the presence of collinearity in a discriminant analysis and a new concept was introduced, i.e. a bias on the the data matrix using SVD - the data matrix as a whole and also an alternative where the bias through SVD is applied on the class groups separately.

A further review was given of outlier/influential observation detection in discriminant analysis.

In section 5.5 a new concept was introduced, viz. the use of the singular value decomposition and a principal components analysis which according to the hypothesis should take care of some of these above-mentioned problems. This new technique was then applied on two data sets and the results were supported by the 3-dimensional biplot displays.

In section 5.7 comparisons were shown to other recent and not so recent developments together with brief descriptions of some of these techniques. Investigations into these developments are still on.

## 5.9 EXTENSIONS

The uses of the SVD procedures can apparently be extended to more statistical

procedures. I shall briefly indicate some ways I believe this procedure, which is so easy to apply, can be put to use:

1. Canonical correlation and canonical variables - see e.g. Johnson and Wichern (1982):

$$\text{If } \underset{\sim}{X} = \begin{pmatrix} \underset{\sim}{X}^{(1)} \\ \underset{\sim}{X}^{(2)} \end{pmatrix}, \qquad \begin{aligned} \underset{\sim}{X}^{(1)} &= U^{(1)} D^{(1)}_{a_1} V^{(1)} : m\times 1 \\ \underset{\sim}{X}^{(2)} &= U^{(2)} D^{(2)}_{a_2} V^{(2)} : s\times 1 \end{aligned}$$

Let $U = \underset{\sim}{a}'\underset{\sim}{X}^{(1)}$ and $V = \underset{\sim}{b}'\underset{\sim}{X}^{(2)}$ then

$$r_{U,V} = \frac{\underset{\sim}{a}' S_{\underset{\sim}{X}^{(1)}\underset{\sim}{X}^{(2)}} \underset{\sim}{b}}{\sqrt{\underset{\sim}{a}' S_{\underset{\sim}{X}^{(1)}\underset{\sim}{X}^{(1)}} \underset{\sim}{a}}\ \sqrt{\underset{\sim}{b}' S_{\underset{\sim}{X}^{(2)}\underset{\sim}{X}^{(2)}} \underset{\sim}{b}}}$$

The first pair of sample canonical covariates is the pair of linear combinations $U_1 = \underset{\sim}{a}_1'\underset{\sim}{X}^{(1)}$; $V_1 = \underset{\sim}{b}_1'\underset{\sim}{X}^{(2)}$. Note that $U_1$ and $V_1$ have sample variances of 1 and they maximize $r_{U,V}$ so that $\hat{\rho}_1 = r_{U_1,V_1}$ is known as the estimator of the first canonical correlation. This procedure is repeated so that if $U_j = \underset{\sim}{a}_j'\underset{\sim}{X}^{(1)}$ and $V_j = \underset{\sim}{b}_j'\underset{\sim}{X}^{(2)}$ with similar characteristics as mentioned above, but with the restriction that $r_{U_j,V_j}$ is maximized using the linear combinations uncorrelated to those used in previous calculations. In general $\hat{\rho}_j$ can be determined as solutions to the equation

$$\left| S_{\underset{\sim}{X}^{(1)}\underset{\sim}{X}^{(2)}} S^{-1}_{\underset{\sim}{X}^{(2)}\underset{\sim}{X}^{(2)}} S_{\underset{\sim}{X}^{(2)}\underset{\sim}{X}^{(1)}} - \hat{\rho}^2 S_{\underset{\sim}{X}^{(1)}\underset{\sim}{X}^{(1)}} \right| = 0$$

with $\hat{p}_1 \geq \ldots \geq \hat{\rho}_j \geq 0$. The coefficient vectors $\underline{a}_1$, $\underline{b}_1$, $\underline{a}_2$, $\underline{b}_2$,...,$\underline{a}_j$, $\underline{b}_j$ are determined by

$$\begin{bmatrix} -\hat{\rho}_j S_{\underline{X}(1)\underline{X}(1)} & S_{\underline{X}(1)\underline{X}(2)} \\ S_{\underline{X}(2)\underline{X}(1)} & -\hat{\rho}_j S_{\underline{X}(2)\underline{X}(2)} \end{bmatrix} \begin{bmatrix} \underline{a}_j \\ \underline{b}_j \end{bmatrix} = \underline{0}$$

The SVD can be applied to the last two equations. In both cases outlying rows can be down-weighted in order to diminish extensive influence by specific rows.

2. Factor analysis - see e.g. Johnson and Wichern (1982):

Given the ortohogonal factor model with m common factors:

$$\underset{(p\times 1)}{\underline{X}} = \underset{(p\times 1)}{\underline{U}} + \underset{(p\times m)(m\times 1)}{\underline{LF}} + \underset{(p\times 1)}{\underline{\varepsilon}}$$

with the normal assumptions that go with factor analysis.

Now $\Sigma = LL' + \Psi$ where $\Psi = \text{cov}(\underline{\varepsilon})$ = diagonal. If the off-diagonal elements of S are small or those of the sample correlation matrix R are essentially zero, then the variables are not related and there are no common unknown factors so that a factor analysis will lead to nothing.

By applying the SVD and principal components analysis to X it can be established whether there are such uncorrelated variables - even if the

inherent relationship is non-linear in nature. Those variables that are uncorrelated can then be removed from the factor analysis as they may influence the analysis of the correlated variables badly.

3. Manova - Johnson and Wichern (1982):

The two-way fixed effects model for a vector response consisting of p components is

$$\underline{X}_{\ell kr} = \underline{\mu} + \underline{\tau}_{\ell} + \underline{\beta}_{k} + \underline{\gamma}_{\ell k} + \underline{\varepsilon}_{\ell kr}, \quad \ell = 1,\ldots,g; \quad k = 1,\ldots,b; \quad r = 1,\ldots,n$$

with the general assumptions: $\sum_{\ell}\underline{\tau}_{\ell} = \sum_{k}\underline{\beta}_{k} = \sum_{\ell}\underline{\gamma}_{\ell k} = \sum_{k}\underline{\gamma}_{\ell k} = \underline{0}$, $\underline{\varepsilon}_{\ell kr} \sim n_p(0,\Sigma)$ so that we have p measurements replicated n times.

To simplify the analysis use the SVD principal components analysis to investigate each of the $\underline{X}_{\ell kr}$ for r = 1,...,n in order to find outliers/influential observations which may be removed creating a smaller and more manageable total matrix with less rows. The original matrix $\left[\underline{X}_{\ell kr}\right]_{g\times b\times n}$ could also be investigated and diminished using the above mentioned technique, but as we have seen in section 5.6 the situation where a typical classification may occur may lead to the results being not easily analyzed, so a separate matrix analysis approach should be more appropriate.

I believe that there are more situations where the SVD principal components analysis may be useful and would suggest an investigation of the possibilities mentioned in this section as well as other situations which may come to the mind

of the reader.

In the mean time research is still carried out with respect to the application of the penalty function as mentioned in chapter 2 as well the different weighting techniques which were proposed in the same chapter. Research is also being done on the regularization idea of Friedman in different estimation areas as well as the techniques with respect to breakdown points proposed by Rousseeuw and Lopulaä. The weighting system of Mallows together with the ideas of De Jongh et al are investigated in the sense of making it more applicable to the non-linear regression model - e.g. applying the weight directly to the loss function and not to be vector of observations.

Another subject which asks for attention is the weighting of the Mallows rows. Before I come back to this I want to refer to an article by Chalton and Troskie (1993).

The mixed regression estimator arises when prior information is available in the form

$$\underset{\sim}{r}_{q\times 1} = R_{q\times p}\,\underset{\sim}{\beta}_{p\times 1} + \underset{\sim}{u}_{q\times 1} \qquad 5.9.1$$

when the regression model is given by

$$\underset{\sim}{y}_{n\times 1} = X_{n\times p}\,\underset{\sim}{\beta}_{p\times 1} + \underset{\sim}{e}_{n\times 1} \qquad 5.9.2$$

There is a standard test for their compatibility. Chalton and Troskie show that the test is equivalent to a test for outliers applied to the observations

corresponding to the prior information. For a single outlier the test statistic can be used to obtain a robust regression estimator. If 5.9.1 and 5.9.2 are combined to give the mixed model

$$\begin{pmatrix} y \\ r \end{pmatrix} = \begin{pmatrix} X \\ R \end{pmatrix} \underline{\beta} + \begin{pmatrix} e \\ u \end{pmatrix}$$

or

$$\underline{y}^* = X^* \underline{\beta} + \underline{e}^*$$

and assuming independence of $\underline{e}$ and $\underline{u}$ we have

$$E(\underline{e}^* \underline{e}^{*\prime}) = \sigma^2 \Omega = \sigma^2 \begin{bmatrix} W & 0 \\ 0 & V \end{bmatrix}$$

If $V = k^{-1}I$, $W = I$, let $\underline{x}'_{(i)}$ be the ith row of X and let $R = \underline{x}_{(i)}$. Note that k is the constant of proportionality relating to the variances of $\underline{e}$ and $\underline{u}$.

Chalton et al show that if there is a single outlier corresponding to the ith case, then a robust estimator of $\underline{\beta}$ is given by

$$\hat{\underline{\beta}} = (X'_{(i)}X_{(i)} + k\underline{x}_i x'_{(i)})^{-1}(X'_{(i)}\underline{y} + k x_{(i)} y_i) \qquad 5.9.3$$

with $k = \dfrac{a(1 - h_i)}{t_i^2 - a h_{ii}}$, where a is computed from the F-distribution, $h_i$ is the ith diagonal element of the hat matrix $X(X'X)^{-1}X'$ and $t_i^2$ is the standardized residual.

If k = 1 we have the least squares estimator. This is a very interesting result since if k = 0, the ith observation is deleted. If 0 < k < 1 less weight is given to the ith observation. It seems possible to even consider a weight of k > 1. This means that more weight is given to the ith observation implying that it is an extremely important design point.

The Mallows weights has the drawback that it can in fact down weight important design points.

Olkin (1992) gives a matrix formulation of how deviant an observation can be. In an absolute gem of a derivation he proofs the well-known result that the deviation of any particular observation from the mean is bounded by a multiple of the standard deviations, i.e.

$$(x_j - \bar{x})^2 \leq \frac{(n-1)^2}{n} s^2 \qquad 5.9.4$$

with $\bar{x} = \frac{1}{n}\Sigma x_i$ and $s^2 = \frac{1}{n-1}\Sigma(x_i - \bar{x})^2$.

He extends the above mentioned result to the mean of a subset, i.e.

$$(\bar{x}^{(1)} - \bar{x})^2 \leq \frac{n-k}{nk} \sum_{1}^{n} (x_i - \bar{x})^2 \qquad 5.9.5$$

and the difference between the two means of subsets

$$(\bar{x}^{(1)} - \bar{x}^{(2)})^2 \leq \frac{k+\ell}{k\ell} \sum_{1}^{n} (x_i - \bar{x})^2 \qquad 5.9.6$$

where $\bar{x}^{(1)} = \frac{1}{k}\sum_{1}^{k} x_i$ and $\bar{x}^{(2)} = \frac{1}{\ell}\sum_{k+1}^{k+\ell} x_i$ .

The Mallows weights down weights the full $\underline{x}'_{(i)}$ vector where $\underline{x}'_{(i)}$ is the ith row of X by $w_i$, $0 \leq w_i \leq 1$, where for $w_i = 1$, the weight is given and for $w_i = 0$ the ith observation is omitted.

It may be that only one element in vector $\underline{x}'_{(i)}$ is spurious, say $x_{i\ell}$. Thus we want to down weight only $x_{i\ell}$ and not $\underline{x}'_{(i)}$. From 5.9.5 we get

$$(x_{i\ell} - \bar{x}^{(2)})^2 \leq \frac{1+(n-1)}{1\times(n-1)} \sum_{j=1}^{n} (x_{ij} - \bar{x}_j)^2$$

where $\bar{x}^{(2)} = \sum_{(\ell)} x_{ij}/(n-1)$ with $\sum_{(\ell)}$ the sum without index $\ell$, or

$$\begin{aligned}(x_{i\ell} - \bar{x}^{(2)})^2 &\leq \frac{n}{n-1}\Sigma(x_{ij} - \bar{x}_j)^2 \\ &\leq n\,S^2\end{aligned}$$

As $(n-1)S^2_{(\ell)}$ must be smaller than $nS^2$ the first can serve as a means to the solution of our upper bound. This means that to find the upper bound the quadratic equation $(x_{i\ell} - \bar{x}^{(2)})^2 = (n-1)S^2_{(\ell)}$ has to be solved. The spurious element only is then replaced by this upper bound or a multiple of this upper bound, say $dx_{i\ell}$ where $0 \leq d < 1$. Note that the final upper bound is computed excluding the spurious observation. I believe that this approach will not only increase the usefulness of the Mallows approach, but will also make it more appropriate for the non-linear situation. This aspect is still under investigation.

---

# APPENDIX A

## TENSOR ANALYSIS AND CURVATURE IN STATISTICS

### A.1 INTRODUCTION

As the usual notation being used in statistics can become a very cumbersome burden when dealing with nonlinear regression, the tendency to use tensor analysis together with the Einsteinian notation has become more and more popular. The simultaneous deletion of the vector notation however, creates problems in so far as legibility is concerned. In this section the vector notation was used where it seemed appropriate in order to emphasize the meaning of some of the expressions. A further important point is that because of the use of the Einsteinian notation, it is necessary to change the indicators for vector components from subscripts to superscripts. A similar way of representing this information, i.e. a combination of the tensor approach and the usual vector-matrix notation has - to the best of my knowledge - not been attempted before. Amari (1982) used the geometrical approach in his analysis of curvature in statistics.



### A.2 ESTIMATION

Let

$\underset{\sim}{\theta} = \{\theta^1, \theta^2, \ldots, \theta^p\} = (\theta^i) = [\theta^i]_{p\times 1}$

$S = \{p(\underset{\sim}{x}, \underset{\sim}{\theta})\}$ the statistical model, $p(\underset{\sim}{x}, \underset{\sim}{\theta})$ being the p.d.f.

$\underset{\sim}{\theta} \in U$ an open subset of $\mathbb{R}^n$.

Note that any allowable coordinate system may be used to analyze the geometric properties of S. However, there are some systems which are more convenient to deal with than others for a specific model in S, and one should make use of such advantages.

Note the following conditions for the geometrical theory to be valid:

a) all $p(\underline{x},\underline{\theta})$'s are such that $p(\underline{x},\underline{\theta}) > 0 \quad \forall \underline{x} \in X$,

b) if $\ell(\underline{x},\underline{\theta}) = \log(p(\underline{x},\underline{\theta}))$, then for every fixed $\underline{\theta}$, p functions of the form $\frac{\partial}{\partial\theta^i}\ell(\underline{x},\underline{\theta})$ are linearly independent i.e. for $i = 1,\ldots,p$, A.2.1

c) the moments of $\frac{\partial}{\partial\theta^i}\ell(\underline{x},\underline{\theta})$ exist up to necessary orders,

d) for the measure P being used and any appropriate $f(\underline{x},\underline{\theta})$

$$\frac{\partial}{\partial\theta^i}\int f(\underline{x},\underline{\theta})dP = \int \frac{\partial}{\partial\theta^i} f(\underline{x},\underline{\theta})dP \text{ for any } f(\underline{x},\underline{\theta}). \qquad \text{A.2.2}$$

The tangent space at any point p of a manifold S can be seen as a vector space obtained using the linearization of S around p. Let c, i.e. the smooth curve $c = c(t) \in S$, form the continuous mapping into S from $t \in [a,b] \in \mathbb{R}^1$. If $\underline{\theta} = \underline{\phi}(p)$, then $\underline{\theta}(t)$ forms the parametric representation of the curve c.

Formally the tangent space can be defined as follows:

Consider F, the set of all smooth real functions on S. If it is given that $\underline{\theta}$ is the coordinate system, then $f = f(\underline{\theta}) \in F$ is a smooth function in $\underline{\theta}$. If $c = c(t) = \underline{\theta}(t)$ is a smooth curve and $f \in F$ then function $f \cdot c$: $[a,b] \rightarrow R^1$, can be written as $f(\underline{\theta}(t))$.

Define the derivative of this function as

$$Cf = \frac{d(f.c)}{dt} = \frac{df(\underline{\theta}(t))}{dt} = \sum_{i=1}^{p} \frac{d\theta^i}{dt} \frac{\partial}{\partial \theta^i} f \qquad \text{A.2.3}$$

which is the derivative of f along the curve or in the direction of the tangent of c. C is known as the directional derivative operator and such an operator C is associated with each curve, where C depends on the tangent vector $\frac{d\theta^i}{dt}$ of the curve c.

At each point $c(t_0)$ on the curve the operator C must satisfy the following conditions:

a) C is a linear mapping from F to R

b) $C(fg) = (Cf)g + f(Cg)$ for $f,g \in F$. A.2.4

Furthermore, a mapping C satisfying the above-mentioned conditions is always derived as the directional derivative operator of a curve.

If the set of mappings C forming an n-dimensional vector space in S is sufficiently smooth, i.e. is $C^\infty$-manifold, it is called the tangent space $T_0$ of S at $p_0$.

Define now the tangent vector $\underline{c}_1$ of $c_1$, the curve along which only $\theta^1$ changes and the other coordinates are fixed, as the partial derivative to $\theta^1$, i.e.

$$\begin{aligned}\underline{c}_1 f &= \frac{d}{dt} f(\underline{\theta}_1(t)) \\ &= \frac{d}{dt} f(\theta_0^1 + t,\ \theta_0^2,\ \theta_0^3,\ldots,\theta_0^p) \\ &= \frac{\partial}{\partial\theta^1} f \end{aligned} \qquad \text{A.2.5}$$

where $\underline{\theta}_0 = (\theta_0^1,\ \theta_0^2,\ \theta_0^3,\ldots,\theta_0^p)$ at point $p_0$ and $c_i$, $i = 1,\ldots,p$, are known as the coordinate curves, i.e. the tangent vector $C_1$ may be denoted by $\frac{\partial}{\partial\theta^1}$ or just $\underline{\partial}_1$. Note that the p vectors $\underline{\partial}_i$ are linearly independent forming a natural basis for $T_0$ associated with the coordinate system $\underline{\theta}$.

It follows therefore that any tangent vector $\underline{a} \in T_0$ can be represented as a linear combination of $\underline{\partial}_i$, i.e.

$$\underline{a} = \sum_{i=1}^{p} a^i \underline{\partial}_i \qquad \text{A.2.6}$$

$a^i$ being the components of $\underline{a}$ with respect to the natural basis $\{\underline{\partial}_i\}$ so that

$$\underline{a} = a^i \underline{\partial}_i \qquad \text{A.2.7}$$

using the Einsteinian summation convention.

Thus if the "$\cdot$" indicates $\frac{d}{dt}$ then $\dot{\underline{\theta}} f = \frac{d}{dt} f(\underline{\theta}(t)) = \dot{\theta}\frac{\partial}{\partial\theta^1} f$ so that the tangent vector of a curve $\underline{\theta}(t)$ in the coordinate expression is given by $\dot{\underline{\theta}} = \dot{\theta}^i \underline{\partial}_i$, i.e.

$\dot{\theta}^i$ denotes the ith component of the tangent vector $\dot{\underline{\theta}}$ of a curve in $\underline{\theta}(t)$.

In statistics we have the manifold $S = \{f(\underline{x},\underline{\theta})\}$ with

$$\ell(\underline{x},\underline{\theta}) = \log f(\underline{x},\underline{\theta}) \qquad \text{A.2.8}$$

and the partial derivatives $\underline{\partial}_i \ell(\underline{x},\underline{\theta})$ being linearly independent functions in $\underline{x}$ for every fixed $\underline{\theta}$. The following p-dimensional vector space is then spaced by p functions $\underline{\partial}_i \ell(\underline{x},\underline{\theta})$ in $\underline{x}$.

$$T_{\underline{\theta}}^{(1)} = \{\underline{a}(\underline{x}) | \underline{a}(\underline{x}) = a^i \underline{\partial}_i \ell(\underline{x},\underline{\theta})\} \qquad \text{A.2.9}$$

where $a^i$, $i = 1,\ldots,p$ are the components of $\underline{a}(\underline{x})$ with respect to the basis $\underline{\partial}_i\, l(\underline{x},\underline{\theta})$. Since $\underline{x}$ is random , $T_{\underline{\theta}}^{(1)}$ is the linear space spanned by $\underline{\partial}_i \ell(\underline{x},\underline{\theta})$. Note further the natural isomorphism between the two vector spaces $T_{\underline{\theta}}$ and $T_{\underline{\theta}}^{(1)}$ i.e.

$$\underline{\partial}_i \in T_{\underline{\theta}} \longleftrightarrow \underline{\partial}_i\, \ell(\underline{x},\underline{\theta}) \in T_{\underline{\theta}}^{(1)} \qquad \text{A.2.10}$$

It is obvious that $T_{\underline{\theta}}$ and $T_{\underline{\theta}}^{(1)}$ are identical in the sense that the former is the differentiation operator representation of the tangent space while $T_{\underline{\theta}}^{(1)}$ is the random variable representation of the same tangent space and is called the 1-representation of the tangent space.

It is of interest to consider the expected value of $h(\underline{x})$ with respect to $f(\underline{x},\underline{\theta})$,

i.e.

$$E[h(\underline{x})] = \int h(\underline{x},\underline{\theta})f(\underline{x},\underline{\theta})dF$$

Differentiate $\int f(\underline{x},\underline{\theta})dF = 1$ with respect to $\theta^i$, i.e.

$$0 = \underline{\partial}_i \int f(\underline{x},\underline{\theta})dF = \int \underline{\partial}_i f(\underline{x},\underline{\theta})dF$$

$$= \int f(\underline{x},\underline{\theta})\underline{\partial}_i \ell(\underline{x},\underline{\theta})dP = E[\underline{\partial}_i \ell(\underline{x},\underline{\theta})]$$

Hence for any $\underline{a}(\underline{x})$ belonging to $T^{(1)}_{\underline{\theta}}$

$$E[\underline{a}(\underline{x})] = \underline{0} \qquad \text{A.2.11}$$

## A.3 TRANSFORMATIONS OR COORDINATE CHANGES

Consider now the two coordinate systems

$$\underline{\theta} = (\theta^1,\ldots,\theta^p) = (\theta^i)$$
$$\underline{\xi} = (\xi^1,\ldots,\xi^p) = (\xi^a)$$

both specifying a distribution in S. There is a diffeomorphism between $\underline{\theta} = \underline{\theta}(\underline{\xi})$ and $\underline{\xi} = \underline{\xi}(\underline{\theta})$ or in component form

$$\xi^a = \xi^a(\theta^1,\ldots,\theta^p), \qquad \theta^i = \theta^i(\xi^1,\ldots,\xi^p)$$

We use $\alpha$, i = 1,...,p, i.e. Greek indices for the $\underline{\xi}$ coordinate system and Arabic for the $\underline{\theta}$ coordinate system. The Jacobian matrices then become

$$[b_i^{\alpha}(\underline{\theta})] = \left[ \frac{\partial \xi^{\alpha}}{\partial \theta^i} \right], \quad [b_{\alpha}^{i}(\underline{\xi})] = \left[ \frac{\partial \theta^i}{\partial \xi^{\alpha}} \right] \qquad \text{A.3.1}$$

Consider the identity $\underline{\theta}[\underline{\xi}(\underline{\theta})] = \underline{\theta}$ or componentwise

$$\theta^i[\xi^1(\underline{\theta}),\ldots,\xi^p(\underline{\theta})] = \theta^i$$

Differentiate this identity now with respect to $\theta^j$

$$\left[ \frac{\partial \theta^i}{\partial \xi^{\alpha}} \right] \left[ \frac{\partial \xi^{\alpha}}{\partial \theta^j} \right] = [b_{\alpha}^{i}(\underline{\xi})][b_j^{\alpha}(\underline{\theta})] = [\delta_j^i] \qquad \text{A.3.2}$$

$\delta_j^i$ being the Kronecker delta. In a similar way take the identity $\underline{\xi}[\underline{\theta}(\underline{\xi})] = \underline{\xi}$, or componentwise

$$\xi^{\alpha}[\theta^1(\underline{\xi}),\ldots,\theta^p(\underline{\xi})] = \xi^{\alpha}$$

and by differentiation with respect to $\xi^{\beta}$

$$\left[ \frac{\partial \xi^{\alpha}}{\partial \theta^i} \right] \left[ \frac{\partial \theta^i}{\partial \xi^{\beta}} \right] = [b_i^{\alpha}(\underline{\theta})][b_{\beta}^{i}(\underline{\theta})] = [\delta_{\beta}^{\alpha}] \qquad \text{A.3.3}$$

i.e. the two Jacobian matrices $[b_i^{\alpha}]$ and $[b_{\alpha}^{i}]$ are mutually inverse matrices. If now $\{\underline{\partial}_i\}$ and $\{\underline{\partial}_{\alpha}\}$ are the natural bases of the tangent space with respect to $\underline{\theta}$ and $\underline{\xi}$ respectively then

$$\underline{\partial}_a = b^i_a \underline{\partial}_i, \quad \underline{\partial}_i = b^a_i \underline{\partial}_a$$

i.e. the vector $\underline{a}$ is

$$\underline{a} = a^i \underline{\partial}_i = a^a \underline{\partial}_a$$

i.e.

$$a^i \underline{\partial}_i = a^a b^i_a \underline{\partial}_i$$

or

$$a^i = a^a b^i_a \qquad \text{A.3.5}$$

and similarly

$$a^a = a^i b^a_i \qquad \text{A.3.6}$$

showing the component change given by a coordinate transformation.

In a similar way the 1-representation $\underline{a}(\underline{x})$ of $\underline{a}$ is invariant for any coordinate systems and only the components change in a contravariant way as the basis changes, i.e.

$$\underline{a}(\underline{x}) = a^i \underline{\partial}_i \ell(\underline{x}, \underline{\theta}) = a^a \underline{\partial}_a \ell(\underline{x}, \underline{\xi}) \qquad \text{A.3.7}$$

## A.4 FISHER INFORMATION IN A RIEMANNIAN SPACE

If one considers now the manifold S with two tangent 1-representation vectors $\underline{a}(\underline{x})$ and $\underline{b}(\underline{x}) \in T_{\underline{\theta}}$, the tangent space, then the inner product is defined as

$$\langle \underline{a}, \underline{b} \rangle = E[\underline{a}(\underline{x})' \underline{b}(\underline{x})] \qquad \text{A.4.1}$$

i.e. the manifold S can be considered a Riemannian space. The inner product

therefore is cov[$\underline{a}(\underline{x})$, $\underline{b}(\underline{x})$], because $E[\underline{a}(\underline{x})] = E[\underline{b}(\underline{x})] = 0$.

Use the two basis vectors $\underline{\partial}_i$ and $\underline{\partial}_j$ and take their inner product in a similar way; then one obtains the metric tensor

$$g_{ij}(\underline{\theta}) = \langle \underline{\partial}_i, \underline{\partial}_j \rangle = E[\underline{\partial}_i \ell(\underline{x}, \underline{\theta}) \underline{\partial}_j \ell(\underline{x}, \underline{\theta})]. \qquad \text{A.4.2}$$

In component form:

$$\langle \underline{a}, \underline{b} \rangle = \langle a^i \underline{\partial}_i, b^j \underline{\partial}_j \rangle = a^i b^j g_{ij} \qquad \text{A.4.3}$$

and for the metric tensor $g_{\alpha\beta}$ in the system $\xi = (\xi^\alpha)$:

$$\begin{aligned} g_{\alpha\beta} &= \langle \underline{\partial}_\alpha, \underline{\partial}_\beta \rangle = \langle b^i_\alpha \underline{\partial}_i, b^j_\beta \underline{\partial}_j \rangle \\ &= b^i_\alpha b^j_\beta \langle \underline{\partial}_i, \underline{\partial}_j \rangle \\ &= b^i_\alpha b^j_\beta g_{ij} \end{aligned} \qquad \text{A.4.4}$$

As the $p^2$ components $g_{ij}$ from the coordinate system $\underline{\theta} = (\theta^i)$ are expressed by the factor $b^i_\alpha b^j_\beta$ as $p^2$ components $g_{\alpha\beta}$ in another coordinate system $\xi = (\xi^\alpha)$, the metric tensor is called a covariant tensor of order 2.

If the two vectors $\underline{a}$ and $\underline{b}$ in the tangent space are orthogonal then $\langle \underline{a}, \underline{b} \rangle = a^i b^j g_{ij} = 0$, i.e. the 1-representations $\underline{a}(\underline{x})$ and $\underline{b}(\underline{x})$ are uncorrelated with zero covariance. Similarly two curves $\theta_1(t)$ and $\theta_2(t)$ passing through a point $\underline{\theta}_0 = \underline{\theta}_1(0) = \underline{\theta}_2(0)$ are said to be orthogonal at this point, $t = 0$, when the tangent

vectors $\dot{\underline{\theta}}_1(0)$ and $\dot{\underline{\theta}}_2(0)$ are orthogonal at this point, i.e.

$$\langle \dot{\theta}_1, \dot{\theta}_2 \rangle = \dot{\theta}_1^i \dot{\theta}_2^j g_{ij} = 0 \qquad \text{A.4.5}$$

or

$$\text{cov}[\dot{\ell}\{\underline{x}, \theta_1(t)\}, \dot{\ell}\{\underline{x}, \theta_2(t)\}] = 0 \qquad \text{A.4.6}$$

at $t = 0$ and length $|\underline{a}|$ is defined by

$$|\underline{a}|^2 = \langle \underline{a}, \underline{a} \rangle = a^i a^j g_{ij} \qquad \text{A.4.7}$$

which is the variance of the 1-presentation $\underline{a}(\underline{x})$

$$|\underline{a}|^2 = E[\{\underline{a}(\underline{x})\}^2] \qquad \text{A.4.8}$$

Denote by $\overrightarrow{pp'} = d\theta^i \underline{\partial}_i$ the infinitesemal vector in $T_{\underline{\theta}}$ where p and p′ are infinitesimal close points with coordinates $\underline{\theta}$ and $\underline{\theta} + d\underline{\theta}$. Then

$$\begin{aligned} d^2(p, p') &= d^2(p(\underline{x}, \underline{\theta}), p(\underline{x}, \underline{\theta} + d\underline{\theta})) \\ &= (\text{distance between p and p}')^2 \\ &= |\overrightarrow{pp'}|^2 \\ &= \langle \overrightarrow{pp'}, \overrightarrow{pp'} \rangle \\ &= g_{ij} d\theta^i d\theta^j \\ &= g_{ij} \dot{\theta}^i \dot{\theta}^j dt^2 \qquad [\dot{\theta}^i = \frac{d\theta^i}{dt}] \end{aligned} \qquad \text{A.4.9}$$

where $[g_{ij}]$ is the Fisher information matrix, such that if $[g^{ij}]$ is its inverse, then

$$[g^{ij}][g_{jk}] = [\delta^i_k] \qquad \text{A.4.10}$$

Define now the unbiased estimators of the parameters $\underline{\theta}$ based on observation $\underline{x}$ from the true distribution $f(\underline{x},\underline{\theta}) = p(\underline{x}, \underline{\theta})$ as $E[\hat{\underline{\theta}}] = \underline{\theta}$. Then according to Cramér-Rao the covariance of any unbiased estimator $\hat{\underline{\theta}} = [\hat{\theta}^i]$ is bounded by the inverse of the Fisher information matrix, i.e.

$$\text{cov}[\hat{\theta}^i, \hat{\theta}^j] \geq [g^{ij}]$$

i.e.

$$|\text{cov}[\hat{\theta}^i, \hat{\theta}^j] - [g^{ij}]| \geq 0 \qquad \text{A.4.11}$$

see Kendall and Stuart (1969).

The distance s between two points $\underline{\theta}_0 = \underline{\theta}(t_0)$ and $\underline{\theta}_1 = \underline{\theta}(t_1)$ on $c:\underline{\theta}(t)$ a smooth curve is then

$$s = \int ds = \int_{t_0}^{t_1} \sqrt{g_{ij}\dot{\theta}^i\dot{\theta}^j}\, dt \qquad \text{A.4.12}$$

The Riemannian geodesic is then the curve among all curves connecting $\underline{\theta}_0$ and $\underline{\theta}_1$ along which the minimum of s is obtained. The Riemannian distance between two points therefore is the distance along the Riemannian geodesic.

In order to be able to calculate the Fisher information matrix (or metric tensor) easily, remember that

$$[g_{ij}(\underline{\theta})] = -[E(\partial_i\partial_j \ell(\underline{x}, \underline{\theta}))] \qquad \text{A.4.13}$$

i.e.

$$[g_{ij}(\underline{\theta})] = -E\left[ \frac{1}{p(\underline{x}, \underline{\theta})} \{\underline{\partial}_i \underline{\partial}_j p(\underline{x}, \underline{\theta}) - p(\underline{x},\underline{\theta})\underline{\partial}_i \ell(\underline{x}, \underline{\theta})\underline{\partial}_j \ell(\underline{x}, \underline{\theta})\}\right] \quad \text{A.4.14}$$

When $\underline{\theta} = \hat{\underline{\theta}} \equiv$ the m.l.e., it satisfies $\underline{\partial}_i \ell(\underline{x}, \hat{\underline{\theta}}) = \underline{0}$ so that $\ell(\underline{x}, \hat{\underline{\theta}})$ is maximized with

$$\ell(\underline{x}, \underline{\theta}) = \ell(\underline{x}, \hat{\underline{\theta}}) + \frac{1}{2}\underline{\partial}_i\underline{\partial}_j \ell(\underline{x}, \hat{\underline{\theta}})(\theta^i - \hat{\theta}^i)(\theta^j - \hat{\theta}^j) \quad \text{A.4.15}$$

We know that $-\underline{\partial}_i\underline{\partial}_j \ell(\underline{x}, \hat{\underline{\theta}})$ shows at what rate $\ell(\underline{x}, \hat{\underline{\theta}})$ is changing at $\hat{\underline{\theta}}$, i.e. the Fisher information is the negative of the expectation of this second derivative of $\ell(\underline{x}, \underline{\theta})$.

## A.5 THE AFFINE CONNECTION AND CURVATURE

Consider $\underline{\theta} \in S$ and $\underline{a}(\underline{\theta}) \in T_{\underline{\theta}}$ with $A = \{\underline{a}(\underline{\theta}) | \underline{\theta} \in S\} = a^i(\underline{\theta})\underline{\partial}_j(\underline{\theta})$ the appropriate vector field as the mapping from S to $T_{\underline{\theta}}$. Note that the vector $\underline{\partial}_i = \underline{\partial}_i(\underline{\theta}) \in T_{\underline{\theta}}$ is a mapping from $T_{\underline{\theta}}$ to $\underline{\theta} \in S$. It can further be said that if the components $a^i(\underline{\theta})$ are smooth functions in $\underline{\theta}$ then the basis vector field $\underline{\partial}_i$ is a smooth vector field. We use the label $T = T(S)$ for the set of all smooth vector fields of S.

Let $\underline{\theta} = \underline{\theta} + d\underline{\theta}$, $d\underline{\theta}$ such that $\underline{\theta}$ and $\underline{\theta}'$ are very close, with corresponding tangent spaces $T_{\underline{\theta}}$ and $T_{\underline{\theta}'} = T_{\underline{\theta}+d\underline{\theta}}$. The linear mapping $m: T_{\underline{\theta} + d\underline{\theta}} \rightarrow T_{\underline{\theta}}$ depends on $d\underline{\theta}$. Similarly $\underline{\partial}_j' = \underline{\partial}_j(\underline{\theta} + d\underline{\theta}) \in T_{\underline{\theta} + d\underline{\theta}} \rightarrow m(\underline{\partial}_j')$ close to $\underline{\partial}_j(\underline{\theta})$. Define

$$\Delta\underline{\partial}_j = \mathbf{m}(\underline{\partial}_j') - \underline{\partial}_j(\underline{\theta}) \in T_{\underline{\theta}}$$

$$= \mathbf{m}[\underline{\partial}_j(\underline{\theta} + d\underline{\theta})] - \underline{\partial}_j(\underline{\theta}) \in T_{\underline{\theta}}$$

$$\simeq d\theta^i \Gamma^k_{ij}(\underline{\theta})\underline{\partial}_k \qquad \text{A.5.1}$$

i.e.

$$\mathbf{m}(\underline{\partial}_j') = \underline{\partial}_j(\underline{\theta}) + \Delta\underline{\partial}_j$$

$$= \underline{\partial}_j(\underline{\theta}) + d\theta^i \Gamma^k_{ij}(\underline{\theta})\underline{\partial}_k \qquad \text{A.5.2}$$

or

$$a^i(\underline{\theta})\mathbf{m}(\underline{\partial}_i') = (a^k(\underline{\theta}) + d\theta^i \Gamma^k_{ij}(\underline{\theta})a^j(\underline{\theta}))\underline{\partial}_k \in T_{\underline{\theta}} \qquad \text{A.5.3}$$

The affine correspondence between $T_{\underline{\theta}}$ and $T_{\underline{\theta}+d\underline{\theta}}$ through $\mathbf{m}$ is obtained by considering the image of the origin of $T_{\underline{\theta}+d\underline{\theta}}$ to be the point $d\theta^i\underline{\partial}_i$ in $T_{\underline{\theta}}$. Therefore obtain the image of $a^i\underline{\partial}_i \in T_{\underline{\theta}+d\underline{\theta}}$ in $T_{\underline{\theta}}$. The following vector is obtained

$$(a^k + d\theta^k + d\theta^i \Gamma^k_{ij}(\underline{\theta})a^j)\underline{\partial}_k$$

As there are now p "equations" with p coefficients each for each of p basis vectors, there are altogether $p^3$ coefficients of the affine connection, viz $\Gamma^k_{ij}(\underline{\theta})$.

The rate of the intrinsic change of $\underline{\partial}_j(\underline{\theta})$ is indicated by

$$\nabla_{\underline{\partial}_i}\underline{\partial}_j = \frac{\partial(\nabla\underline{\partial}_j)}{\partial(d\theta^i)}$$

$$= \Gamma^k_{ij}(\underline{\theta})\underline{\partial}_k \qquad \text{A.5.4}$$

where $\nabla\underline{\partial}_j$ can be seen as the intrinsic change in the j'th basis vector $\underline{\partial}_j(\underline{\theta})$ as the point changes from $\underline{\theta}$ to $\underline{\theta} + d\underline{\theta}$.

The vector field $\nabla_{\underline{\partial}_i}\underline{\partial}_j$ is called the covariant derivative of the vector field $\underline{\partial}_j$ along $\underline{\partial}_i$, i.e. it measures the rate of intrinsic change as $\underline{\theta}$ changes in the direction of $\underline{\partial}_i$.

As $\nabla_{\underline{\partial}_i}\underline{\partial}_j$ can be determined from $\Gamma^k_{ij}(\underline{\theta})$, the inverse is also true.

Take the inner product of the r.h.s. and l.h.s. and the result follows:

$$\langle\nabla_{\underline{\partial}_i}\underline{\partial}_j, \underline{\partial}_m\rangle = \langle\Gamma^k_{ij}(\underline{\theta})\underline{\partial}_k, \underline{\partial}_m\rangle = \Gamma^k_{ij}(\underline{\theta})\langle\underline{\partial}_k, \underline{\partial}_m\rangle$$

$$= \Gamma^k_{ij}(\underline{\theta})g_{km}(\underline{\theta})$$

$$= \Gamma_{ijm}(\underline{\theta}) \qquad \text{A.5.5}$$

i.e.

$$\Gamma^k_{ij}(\underline{\theta}) = g^{km} = \Gamma_{ijm}(\underline{\theta}) \qquad \text{A.5.6}$$

where $[g^{km}]$ is the inverse matrix of $[g_{km}]$. This means that the covariant derivative $\nabla$ is an affine connection on S.

For two vector fields A and B $\in$ T(S) define the covariant derivative of B along A as the vector field $C = \nabla_A B$; $\underline{C}(\underline{\theta}) \in T(\underline{\theta})$ being the rate of intrinsic change in vector field $B(\underline{\theta})$ as the point $\underline{\theta}$ changes in the direction of vector $\underline{a}(\underline{\theta})$. Therefore, given vector fields $A, A', B, B' \in T(S)$ and a smooth scalar function $f: S \rightarrow R$ such that $f(\underline{\theta})\underline{a}(\underline{\theta}) \in T(\underline{\theta})$, i.e. fA is also a vector field.

(a) $\nabla_A(B + B') = \nabla_A B + \nabla_A B'$

(b) $\nabla_{(A + A')} B = \nabla_A B + \nabla_{A'} B$

(c) $\nabla_A(fB) = (Af)B + f\nabla_A B$

(d) $\nabla_{fA}(B) = f\nabla_A B$ with

(e) $Af = a^i(\underline{\theta})\underline{\partial}_i f(\underline{\theta})$ A.5.7

and again

$$\Gamma_{ijk} = \langle \nabla_{\underline{\partial}_i} \underline{\partial}_j, \underline{\partial}_k \rangle \qquad \text{A.5.8}$$

or given $\Gamma_{ijk}(\underline{\theta})$ —

$$\nabla_A(B) = \nabla_A(b^j \underline{\partial}_j)$$

$$= (Ab^j)\underline{\partial}_j + b^j \nabla_A \underline{\partial}_j$$

$$= (Ab^j)\underline{\partial}_j + b^j \nabla_{a^i \underline{\partial}_i} \underline{\partial}_j$$

$$= (a^i \underline{\partial}_j b^j)\underline{\partial}_j + b^j a^i \Gamma^k_{ij} \underline{\partial}_k$$

$$= (a^i \underline{\partial}_j b^k + b^j a^i \Gamma^k_{ij})\underline{\partial}_k \qquad \text{A.5.9}$$

It is obvious that the coefficients of the affine connection are dependent on the coordinate system $\underline{\theta}$. Say we have the coordinate system $\xi = (\xi^a)$ with the Jacobian elements

$$B^i_a = \partial\theta^i/\partial\xi^a \qquad \text{A.5.10}$$

i.e.

$$B^i_a \underline{\partial}_i = \frac{\partial\theta^i}{\partial\xi^a}\frac{\partial}{\partial\theta^i} = \frac{\partial}{\partial\xi^a} = \underline{\partial}_a = \partial_a \qquad \text{A.5.11}$$

then

$$\Gamma_{a\beta\gamma}(\xi) = \langle \nabla_{\underline{\partial}_a}\underline{\partial}_\beta, \underline{\partial}_\gamma\rangle$$

$$= \langle \nabla_{B^i_a\underline{\partial}_i}(B^j_\beta\underline{\partial}_j), B^k_\gamma\underline{\partial}_k\rangle$$

$$= B^i_a B^k_\gamma \langle \nabla_{\underline{\partial}_i} B^j_\beta\underline{\partial}_j, \underline{\partial}_k\rangle$$

$$= B^i_a B^k_\gamma \langle \underline{\partial}_i B^j_\beta\underline{\partial}_j + B^j_\beta\nabla_{\underline{\partial}_i}\underline{\partial}_j, \underline{\partial}_k\rangle$$

$$= B^i_a B^k_\gamma (\langle \underline{\partial}_i B^j_\beta\underline{\partial}_j, \underline{\partial}_k\rangle + \langle B^j_\beta\nabla_{\underline{\partial}_i}\underline{\partial}_j, \underline{\partial}_k\rangle)$$

$$= B^i_a B^k_\gamma (\underline{\partial}_i B^j_\beta\langle\underline{\partial}_j, \underline{\partial}_k\rangle + B^j_\beta\langle\nabla_{\underline{\partial}_i}\underline{\partial}_j, \underline{\partial}_k\rangle)$$

$$= g_{jk}B^k_\gamma\partial_a B^j_\beta + B^i_a B^j_\beta B^k_\gamma \Gamma_{ijk} \qquad \text{A.5.12}$$

The affine connection can be used in the space S of a statistical model in order to be able to study the intrinsic properties of the family of probability distributions. As the natural basis $\underline{\partial}_j(\underline{\theta} + d\underline{\theta})$ of $T_{\underline{\theta} + d\underline{\theta}}$ is represented by

$$\underline{\partial}_j \ell(\underline{x}, \underline{\theta} + d\underline{\theta}) = \underline{\partial}_j \ell(\underline{x}, \underline{\theta}) + \underline{\partial}_i \underline{\partial}_j \ell(\underline{x}, \underline{\theta}) d\theta^i \qquad \text{A.5.13}$$

Now it can be shown that the affine connection with the following coefficients can be obtained,

$$\Gamma^{(\alpha)}_{ijk}(\underline{\theta}) = E[\{\underline{\partial}_i \underline{\partial}_j \ell(\underline{x}, \underline{\theta}) + \frac{1-\alpha}{2} \underline{\partial}_i \ell(\underline{x}, \underline{\theta}) \underline{\partial}_j \ell(\underline{x}, \underline{\theta})\} \underline{\partial}_k \ell(\underline{x}, \underline{\theta})] \qquad \text{A.5.14}$$

and this is called the $\alpha$-connection which, when $\alpha = 1$, reduces to the exponential connection

$$\Gamma^{1}_{ijk}(\underline{\theta}) = \Gamma_{ijk}(\underline{\theta}) = E[\underline{\partial}_i \underline{\partial}_j \ell(\underline{x}, \underline{\theta}) \underline{\partial}_k \ell(\underline{x}, \underline{\theta})] \qquad \text{A.5.15}$$

and when $\alpha = -1$, to the mixture connection

$$\Gamma^{-1}_{ijk}(\underline{\theta}) = \Gamma_{ijk}(\underline{\theta}) = E[\{\underline{\partial}_i \underline{\partial}_j \ell(\underline{x}, \underline{\theta}) + \underline{\partial}_i \ell(\underline{x}, \underline{\theta}) \underline{\partial}_j \ell(\underline{x}, \underline{\theta})\} \underline{\partial}_k \ell(\underline{x}, \underline{\theta})] \qquad \text{A.5.16}$$

At this stage one can define the third-order tensor

$$T_{ijk}(\underline{\theta}) = E[\underline{\partial}_i \ell(\underline{x}, \underline{\theta}) \underline{\partial}_j \ell(\underline{x}, \underline{\theta}) \underline{\partial}_k \ell(\underline{x}, \underline{\theta})] \qquad \text{A.5.17}$$

with cange in components under coordinate transformations as follows

$$T_{\alpha\beta\gamma} = B^{i}_{\alpha} B^{j}_{\beta} B^{k}_{\gamma} T_{ijk} \qquad \text{A.5.18}$$

The $\alpha$-connection can now be written as

$$\Gamma^{(\alpha)}_{ijk} = \Gamma^{(1)}_{ijk} + \frac{1-\alpha}{2} T_{ijk} \qquad \text{A.5.19}$$

where (1) refers to the 1-representation of the tangent space.

The $\alpha$-connection now defines the $\alpha$-straightness in the sense of intrinsic linearity of the model, so that deviation from it is measured by the $\alpha$-curvature.

A special case of the affine connection is a metric. This is the case when $\alpha = 0$ and is known as the Riemannian or Levi-Civita or information connection. In this connection the minimum length curve between two points (geodesic) is given by a straight line and this is the only connection for which the parallel shift of a vector does not change its length - therefore a metric. The Riemannian metric tensor $g_{ij}(\underline{\theta})$ has a connection whose coefficients are given by

$$[i, j; k] = \frac{1}{2}[\underline{\partial}_i g_{jk} + \underline{\partial}_j g_{ik} - \underline{\partial}_k g_{ij}] \qquad \text{A.5.20}$$

which is also called the Christoffel 3-indices symbol.

<u>Some results can now be summarized</u>:

The following mappings are useful:

(a) A covariant tensor of order k, i.e. a multi-linear mapping from k vector fields to the real numbers R:-

$$Q: T(S)\times T(S)\times T(S)\times\ldots\times T(S) \rightarrow R \qquad \text{A.5.21}$$

It is good to remember that when the components of a tensor vanishes in a coordinate system they vanish in any coordinate system. The Riemannian metric is a tensor of order two with components $g_{ij}$. Note that the affine connection is not a tensor, because there is no multilinear mapping involved and its transformation rule is different.

(b) A multilinear mapping from k vector fields to a vector field T(S) is called a tensor of order k + 1, of covariant order k and contravariant order 1:-

$$R: T(S)\times T(S)\times T(S)\times\ldots\times T(S) \rightarrow T(S) \qquad \text{A.5.22}$$

(c) A bilinear mapping from $T(S)\times T(S) \rightarrow T(S)$ using an affine connection. This is a tensor of order three and is known as the torsion.

(d) The trilinear mapping $R: T(S)\times T(S)\times T(S) \rightarrow T(S)$ using an affine connection, i.e. a tensor of order four which is known as the Riemann-Christoffel curvature. For three vector fields A, B, C $\in$ T(S)

$$R(A, B, C) = [\nabla_A, \nabla_B]C - \nabla_{[A,B]}C$$

$$= (\nabla_A\nabla_B - \nabla_B\nabla_A)C - \nabla_{AB-BA}C \qquad \text{A.5.23}$$

Given a coordinate system $\underline{\theta}$ the curvature is represented by the components of a tensor

$$R_{ijkm} = \langle R(\underline{\partial}_i, \underline{\partial}_j, \underline{\partial}_k), \underline{\partial}_m \rangle \qquad \text{A.5.24}$$

being called the Riemann-Christoffel curvature tensor, and

$$R_{ijkm} = (\underline{\partial}_i \Gamma^s_{jk} - \underline{\partial}_j \Gamma^s_{ik}) g_{sm} + (\Gamma_{irm}\Gamma^r_{jk} - \Gamma_{jrm}\Gamma^r_{ik}) \qquad \text{A.5.25}$$

When $R_{ijkm}(\underline{\theta}) = 0$ for any $\underline{\theta}$, or $R(A, B, C) = 0$ for any $A, B, C \in T(S)$ then a space with an affine connection is said to be flat.

If $R_{ijkm} = 0$ at any point in S, there exists a coordinate system $\underline{\theta} = (\theta^i)$ in a torsion free manifold such that $\Gamma_{ijk}(\underline{\theta}) = 0$. From this follows that $\nabla_{\underline{\partial}_i}\underline{\partial}_j = 0$, i.e. the basis vector fields $\underline{\partial}_i$ are parallel vector fields. Such a coordinate system is said to be affine. There exists no affine coordinate systems in S, unless S is curvature free. Not withstanding this there does exist for any point $\underline{\theta}_0$ a coordinate system such that the coefficients of the affine connection and their derivatives vanish at this one point $\underline{\theta}_0$, i.e.

$$\Gamma_{ijk}(\underline{\theta}_0) = \underline{\partial}_m \Gamma_{ijk}(\underline{\theta}_0) = \underline{\partial}_s \underline{\partial}_m \Gamma_{ijk}(\underline{\theta}_0) = 0 \qquad \text{A.5.26}$$

Such a system is called a normal coordinate system at $\underline{\theta}_0$.

In spite of many publications on the subject the statistical implications of Riemannian structures and especially of the Riemann-Christoffel curvature are

not all that clear. Rao (1945) showed that a statistical model forms a Riemannian manifold and the Fisher information matrix is the metric tensor $g_{ij}$.

The articles that set off further investigation into this subject were those by Efron (1975). See also chapter 2, Dawid (1975) and Reed (1975).

Some extensions were also made on the Riemannian approach by Atkinson and Mitchell (1981), Kass (1984) and Skovgaard (1984). It was shown that a covariant stabilizing transformation or reparameterization of a statistical model, i.e. a transformation such that $g_{ij}(\underline{\theta})$ reduces to the unit matrix $[\delta_{ij}]$ for all $\underline{\theta}$ is only possible if the Riemann-Christoffel curvature vanishes.

---

# APPENDIX B

## THE BASIC STRUCTURE DISPLAY OF A DATA MATRIX

### B.1 INTRODUCTION

The basic structure display of a data matrix (BSDM) is also known as the "canonical form" (Eckart and Young, 1936) or the singular "decomposition" (Good, 1969). This display is also known as the "Eckart-Young decomposition" (see Kristof, 1970).

Greenacre (1980), like Green and Carroll (1976) preferred the term "basic structure" in his research paper where he summarised his description of basic structure in his thesis - see Greenacre (1978).

The basic structure of a matrix is the decomposition of the matrix into elements of sample structure with an immediate geometric appeal. Given a matrix A and using its basic structure one can find a least-squares approximation $\hat{A}$ of A with the feature that rank($\hat{A}$) < rank(A). This $\hat{A}$ now provides a graphical display of the original A.



### B.2 BASIC STRUCTURE - GREENACRE, 1980

Any real matrix $A_{n\times m}$ may be expressed in the basic structure as

$$A_{n\times m} = U_{n\times r}\, D_{a_{r\times r}}\, V'_{r\times m} \qquad \text{B.2.1}$$

$$= \sum_{k=1}^{r} a_k \underset{n\times 1}{\underline{u}_k} \underset{1\times m}{\underline{v}_k'} \qquad \text{B.2.2}$$

where $D_a = \text{diag}\ (a_1,\ldots,a_r)$, $a_i > 0$, $i = 1,\ldots,r$; $r \leq \min(n,m) = \text{rank}(A)$ and $U'U = I = V'V$. Call $a_k$ the kth basic value, $\underline{u}_k$ the kth left basic vector and $\underline{v}_k$ the kth right basic vector.

The column vectors $\underline{u}_k$, $k = 1,\ldots,r$ of U form an orthonormal basis for the columns of A and the column vectors $\underline{v}_k$, $k = 1,\ldots,r$ of V form an orthonormal basis for the transposed rows of A. The matrices U and V thus determine the multidimensional subspace in which A is contained. The basic values in $D_a$ determine the "magnitude" of A in each of its r basic dimensions.

In the special case when A is a symmetric matrix, say $A = B_{n\times n}$ with $\text{rank}(B) = r \leq n$ the basic structure of B is

$$B_{n\times n} = U_{n\times r} D_{\lambda_{r\times r}} U'_{r\times n} \qquad \text{B.2.3}$$

$$= \sum_{k=1}^{r} \lambda_k \underline{u}_k \underline{u}_k' \qquad \text{B.2.4}$$

If it is assumed that the basic values are arranged in descending order so that $a_1 \geq a_2 \geq \ldots \geq a_r > 0$ with the basic vectors of U and V correspondingly then the basic structure is uniquely determined so that one can approximate A by $\hat{A}_{[p]}$ where

$$\hat{A}_{[p]} = \sum_{k=1}^{p} a_k \underset{\sim}{u}_k \underset{\sim}{v}_k' \qquad \text{B.2.5}$$

$\hat{A}_{[p]}$ is the n×m matrix formed from the first p (i.e. the largest) basic values and corresponding basic vectors of the matrix A of rank r where $p < r$. $\hat{A}_{[p]}$ is called the "best rank p approximation" of A in the sense that it minimises

$$\|A - A_{[p]}\|^2 \equiv \text{trace}\left[(A - A_{[p]})(A - A_{[p]})'\right] \qquad \text{B.2.6}$$

for all rank p matrices $A_{[p]}$.

In matrix form $\hat{A}_{[p]}$ can be expressed as

$$\hat{A}_{[p]} = U_1 D_{a_1} V_1' \qquad \text{B.2.7}$$

from

$$\begin{aligned} A &= U_1 D_{a_1} V_1' + U_2 D_{a_2} V_2' \\ &= \hat{A}_{[p]} + (A - \hat{A}_{[p]}) \end{aligned} \qquad \text{B.2.8}$$

where

$$U = \underset{p}{\left[U_1\right.} \vdots \underset{r-p}{\left.U_2\right]} \; n$$

$$V = \underset{p}{\left[V_1\right.} \vdots \underset{r-p}{\left.V_2\right]} \; m$$

$$D_a = \begin{bmatrix} D_{a_1} & \cdots & 0 \\ \vdots & & \\ 0 & \cdots & D_{a_2} \end{bmatrix} \begin{matrix} p \\ \\ r-p \end{matrix}$$

$$\begin{matrix} p & & r-p \end{matrix}$$

$\hat{A}_{[p]}$ is called the rank p basic structure of A where A is the rank r basic structure. A measure of the "fit" of $\hat{A}_{[p]}$, the "least squares estimate" (Referring to the matrix norm) to A is given by

$$\tau_{[p]} = \frac{\|\hat{A}_{[p]}\|^2}{\|A\|^2} = \frac{\text{trace}(\hat{A}_{[p]}\hat{A}'_{[p]})}{\text{trace}\,(AA')}$$

$$= \frac{\sum_{k=1}^{p} a_k^2}{\sum_{k=1}^{r} a_k^2} \qquad \text{B.2.9}$$

so that $0 \leq \tau_{[p]} \leq 1$ and the error of approximation is given by

$$1 - \tau_{[p]} = \frac{\|A - \hat{A}_{[p]}\|^2}{\|A\|^2}$$

$$= \frac{\text{trace}\left[(A - \hat{A}_{[p]})(A - \hat{A}_{[p]})'\right]}{\sum_{k=1}^{r} a_k^2}$$

$$= \frac{\sum_{k=p+1}^{r} a_k^2}{\sum_{k=1}^{r} a_k^2} \qquad \text{B.2.10}$$

Computation of the basic structure can be accomplished by the algorithm of Golub and Reinsch (1971) or by using the fact that if

$$A = UD_aV'$$

then

$$\begin{aligned} A'A &= VD_aU'\ UD_aV' \\ &= VD_a^2V' \end{aligned} \qquad \text{B.2.11}$$

which is the eigenstructure of the m×m symmetric matrix A'A with eigenvalues the squared values $a_k^2$, k = 1,...,r and eigen vectors the right basic vectors $\underset{\sim}{v}_k$, k = 1,...,r of A (refer 2.3). If m ≤ n, find the structure in 2.11, i.e. V and $D_a^2$ and therefore $D_a = +\sqrt{D_a^2}$. Then from 2.1 and 2.11

$$\begin{aligned} A &= UD_aV' \\ AV &= UD_a \\ AVD_a^{-1} &= U \\ \text{or} \qquad U &= AVD_a^{-1} \end{aligned} \qquad \text{B.2.12}$$

If m > n one could use AA' for computational purposes.

## B.3 THE GENERALISED BASIC STRUCTURE

In more general terms the basic structure could be determined using the "generalised Fröbenius norm" in stead of the "Fröbenius norm" where the latter is

$$\|A\|^2 = tr[AA'] = \sum_{ij}^{nm} a_{ij}^2$$
$$= \sum_{i=1}^{n} \underline{a}_i' \underline{a}_i \qquad \text{B.3.1}$$

where $\underline{a}_i$ is the ith row vector of A written as a column vector.

Use the generalised Euclidian norm to define $\phi$:

$$\|\underline{a}_i\|_\phi^2 = \underline{a}_i' \phi \underline{a}_i \qquad \text{B.3.2}$$

where $\phi$ is positive definite.

The generalised Fröbenius norm will then be

$$\|A\|_{D_\omega,\phi}^2 \equiv \sum_{i=1}^{n} \omega_i \|\underline{a}_i\|_\phi^2$$
$$= tr(D_\omega A \phi A') \qquad \text{B.3.3}$$

where $\underline{\omega}$ is just a weighting vector.

The solution to our problem of finding the least-squares lower rank approximation of the matrix A, i.e. $\hat{A}_{[p]}$ where $\hat{A}_{[p]}$ is the solution to the expression:

$$\text{Minimise } \|A - A_{[p]}\|_{\Omega,\phi}^2 = tr\left[\Omega(A - A_{[p]})\phi(A - A_{[p]})'\right] \qquad \text{B.3.4}$$

can be determined using the generalised basic structure of A.

The latter can be obtained from the ordinary basic structure of the appropriately transformed matrix, i.e.

$$A_{n\times m} = N_{n\times r}\, D_{a_{r\times r}}\, M'_{r\times m} \qquad \text{B.3.5}$$

where $N'\Omega N = I = M'\phi M$ gives rise to $\hat{A}_{[p]} = N_1 D_{a_1} M'_1$, because if $N = \Omega^{-\frac{1}{2}}U$ and $M = \phi^{-\frac{1}{2}}V$, then the transformed matrix

$$\Omega^{\frac{1}{2}}A\phi^{\frac{1}{2}} = (\Omega^{\frac{1}{2}}N)D_a(M'\phi^{\frac{1}{2}}) \qquad \text{B.3.6}$$

where $(\Omega^{\frac{1}{2}}N)'(\Omega^{\frac{1}{2}}N) = I = (\phi^{\frac{1}{2}}M)'(\phi^{\frac{1}{2}}M)$, i.e. $A = ND_aM'$ where $N'\Omega N = I = M'\phi M$ as stated. This means that $U_1 D a_1 V'_1$ is replaced by $\Omega^{\frac{1}{2}}N_1 D_{a_1} M'_1\phi^{\frac{1}{2}}$ so that the error

$$\|\Omega^{\frac{1}{2}}A\phi^{\frac{1}{2}} - \Omega^{\frac{1}{2}}N_1 D_{a_1} M'_1\phi^{\frac{1}{2}}\|^2 = \mathrm{tr}\left[\Omega(A - N_1 D_{a_1} M'_1)\phi(A - N_1 D_{a_1} M'_1)'\right]$$
$$= \|A - N_1 D_{a_1} M'_1\|^2_{\Omega,\phi} \qquad \text{B.3.7}$$

is minimised.

As a result $N_1 D_{a_1} M'_1$ is implied as the rank p approximation of A in the general norm.

## B.4 BASIC STRUCTURE DISPLAY

A graphical display which approximates the higher dimensional rectangular data matrix can now be obtained using features of the basic structure.

Assume a data matrix which is preprocessed, for example to "centre" the data and call this processed matrix Z. Then find the lower rank p matrix through the generalised basic structure and call this matrix $\hat{Z}_{[p]}$ where

$$\hat{Z}_{[p]} = N_{1_{n\times p}} D_{a_{1_{p\times p}}} M'_{1_{p\times m}} \qquad \text{B.4.1}$$

The object is to represent the rows of $\hat{Z}_{[p]}$ as points in a p-dimensional Euclidian space, i.e. with p axes, so that the between points distances in the display are exactly the between rows distances in the metric $\phi$. These displayed distances are approximations of the true distances in the metric $\phi$ between rows of the original Z.

Let the coordinates of the row points in the display be contained in the rows of matrix $F_{n\times p}$. The linear structure of F is then the set of scalar products, i.e.

$$\begin{aligned} FF' &= \hat{Z}_{[p]}\phi\hat{Z}'_{[p]} \\ &= N_1 D_{a_1} M'_1 \phi M_1 D_{a_1} N'_1 \\ &= (N_1 D_{a_1})(N_1 D_{a_1})' \end{aligned} \qquad \text{B.4.2}$$

so that one can take F to be $N_1 D_{a_1}$.

Let p be equal to 2, then the basic concept of a biplot permits the columns of $\hat{Z}_{[p]}$ to be represented by

$$G = M_1 \qquad \text{B.4.3}$$

i.e.

$$\begin{aligned}\hat{Z}_{[2]} &= N_1 D_{a_1} M_1' \\ &= F M_1' \\ &= F G' \end{aligned} \qquad \text{B.4.4}$$

i.e.

$$\hat{z}_{ij} = \underline{f}_i' \underline{g}_j \qquad \text{B.4.5}$$

where $\underline{f}_i$ and $\underline{g}_j$ are the ith and jth rows of F and G respectively written as column vectors and $\hat{z}_{ij}$ is an approximation of the (i,j)th element of the original Z.

Other factorisations may be possible:

$$\hat{Z}_{[2]} = F G' \text{ where } F = N_1 \text{ and } G = M_1 D_{a_1} \qquad \text{B.4.6}$$

and

$$\hat{Z}_{[2]} = F G' \text{ where } F = N_1 D_{a_1}^{\frac{1}{2}} \text{ and } G = M_1 D_{a_1}^{\frac{1}{2}} \qquad \text{B.4.7}$$

both having the biplot property that $\hat{z}_{ij} = \underline{f}_i' \underline{g}_j$, but with different meanings. Note further that from $Z = N D_a M'$ with $N'\Omega N = I = M'\phi M$ follow that

$$Z = N D_a M'$$

$$Z\phi\mathbf{M} = N D_a \mathbf{M}'\phi\mathbf{M}$$

$$Z\phi\mathbf{M} = N D_a$$

$$Z\phi\mathbf{M}D_a^{-1} = N \qquad \text{B.4.8}$$

and

$$Z = N D_a \mathbf{M}'$$

$$N'\Omega Z = N'\Omega N D_a \mathbf{M}'$$

$$N'\Omega Z = D_a \mathbf{M}'$$

$$D_a^{-1} N'\Omega Z = \mathbf{M}'$$

$$Z'\Omega N D_a^{-1} = \mathbf{M} \qquad \text{B.4.9}$$

so that for example when $\hat{Z}_{[2]} = FG'$ with $F = N_1 D_{a_1}$ and $G = \mathbf{M}_1$ we have

$$N_1 = \hat{Z}_{[2]}\phi\mathbf{M}_1 D_{a_1}^{-1} \quad \text{and} \quad \mathbf{M}_1 = \hat{Z}'_{[2]}\Omega N_1 D_{a_1}^{-1}$$

$$N_1 D_{a_1} = \hat{Z}_{[2]}\phi\mathbf{M}_1 \quad \text{and} \quad \mathbf{M}_1 = \hat{Z}'_{[2]}\Omega N_1 D_{a_1} D_{a_1}^{-2}$$

$$F = \hat{Z}_{[2]}\phi G \qquad G = \hat{Z}'_{[2]}\Omega F D_{a_1}^{-2} \qquad \text{B.4.10}$$

from 4.7.

## B.5 COMPUTATION OF THE COORDINATES

First symmetrise the matrix to be diagonalised by pre-multiplying by $\phi^{\frac{1}{2}}$. Then, if one assumes that $m \leq n$, solve for $\mathbf{M}$ first by setting up the eigen equation

$$\phi^{\frac{1}{2}}(Z'\Omega Z\phi)M = \phi^{\frac{1}{2}}MD_a^2$$

that is

$$(\phi^{\frac{1}{2}}Z'\Omega Z\phi^{\frac{1}{2}})\phi^{\frac{1}{2}}M = \phi^{\frac{1}{2}}MD_a^2 \qquad \text{B.5.1}$$

where $M'\phi M = I$, i.e. $V'V = (\phi^{\frac{1}{2}}M)'\phi^{\frac{1}{2}}M = I$ and then $M = \phi^{-\frac{1}{2}}V$. Use 4.8 to determine the left basic vectors N.

A symmetric argument for the basic vectors N leads to

$$(Z\phi Z'\Omega)N = ND_a^2 \qquad \text{B.5.2}$$

In 5.2 if $\Omega = I$ then $N = S$, the scalar products in the metric $\phi$ of the rows of Z: $s_{ij} = \underline{z}_i'\phi\underline{z}_j$.

If $\Omega = D_w$ then $N = SD_w = [s_{ij}w_j]$, $w_j$ being the weights assigned to the row points.

## B.6 NOTATION

The notation BSDM (Z; $\Omega$, $\phi$; a, b) for the generalised basic structure display of the data matrix summarises the procedure with the following meanings attached:

(i) $\Omega_{n\times n}$ defines a norm on the columns of Z, or alternatively a set of weights on the rows.

(ii) $\phi_{m\times m}$ defines a norm on the rows of Z, or alternatively a set of weights on the columns.

(iii) If $Z = ND_{a}M'$ is the generalised basic structure of Z with $\Omega$ and $\phi$ defined as above, then the coordinate matrices F and G of the row and column points are: $F = N_1 D^{a}_{a_1}$ and $G = M_1 D^{b}_{a_1}$ with $N_1$, $M_1$ and $D_{a_1}$ as defined in earlier sections.

(iv) a+b = 1 indicates a biplot interpretation as in paragraph B.4.

Approximate Euclidian distances between rows of Z which are unweighted would be obtained by having $\phi$ = I and $\Omega$ = I respectively, i.e. BSDM (Z; I, I; 1, 0).

Note the difference between BSDM (Z; I, I; 1, 0) and BSDM (Z; I, I; 1, -) where the first indicates that the column points are going to be plotted with the biplot interpretation between row and column points (since a+b=1). The latter indicates that only row points are going to be displayed.

## B.7 COMPUTATION AND THE BSDM ANALYSIS

Assume $m \leq n$; then the algorithm is as follows:

(i) Read data matrix X

(ii) Transform X to Z

(iii) Perform BSDM (Z; $\phi$, $\Omega$; a, b) with $\phi$, $\Omega$, a and b specified, i.e.

(a) Compute the symmetric matrix

$$\phi^{\frac{1}{2}}Z'\Omega Z\phi^{\frac{1}{2}} = Q \qquad \text{B.7.1}$$

where $\phi$ is diagonal; if not compute $\phi^{\frac{1}{2}}$ using the eigenstructure of $\phi$:

$$\phi = UD_{\lambda}U'$$

$$= UD_{\lambda}^{\frac{1}{2}}U'UD_{\lambda}^{\frac{1}{2}}U' \qquad \text{B.7.2}$$

so that

$$\phi^{\frac{1}{2}} = UD_{\lambda}^{\frac{1}{2}}U' \qquad \text{B.7.3}$$

(b) Determine the eigenstructure of $Q$

$$Q = VD_{\mu}V' \qquad \text{B.7.4}$$

(c) Find F and G in p dimensions:

$$G = \phi^{-\frac{1}{2}}V_1(D_{\mu}^{\frac{1}{2}})^b \qquad \text{B.7.5}$$

and

$$F = Z\phi^{\frac{1}{2}}V_1(D_{\mu_1}^{\frac{1}{2}})^{a-1} \qquad \text{B.7.6}$$

where $D_{\mu_1}$ and $V_1$ contain the largest p eigenvalues of $Q$ and corresponding eigenvectors respectively.

(d) Complete the plotting routine for row and/or column points in selected pairs of dimensions.

## APPENDIX C

### THE SHERMAN-MORRISON-WOODBURY THEOREM (RAO, 1973 P 33):

The theorem for vectors states that

$$(A - \underline{u}\underline{v}')^{-1} = A^{-1} + \frac{A^{-1}\underline{u}\underline{v}'A^{-1}}{1 - \underline{v}'A^{-1}\underline{u}}$$

or

$$(A + a'b)^{-1} = A^{-1} - A^{-1}a'(I + bA^{-1}a')^{-1}\, bA^{-1}$$

is used where the ranks of matrices A, a and b must conform. Let $A = \hat{J}'\hat{J}$, $a = -\hat{\underline{j}}_i'$ and $b = \hat{\underline{j}}_i'$, then

$$\hat{\theta}_{(i)} = \hat{\underline{\theta}} + (\hat{J}'_{(i)}\hat{J}_{(i)})^{-1}\hat{J}'_{(i)}\underline{e}_{(i)} \quad \text{from 3.4.5}$$

$$= \hat{\underline{\theta}} + (\hat{J}'\hat{J} - \hat{\underline{j}}_i\hat{\underline{j}}_i')^{-1}(\hat{J}'\underline{e} - \hat{\underline{j}}_i e_i)$$

$$= \hat{\underline{\theta}} + \left[(\hat{J}'\hat{J})^{-1} + (\hat{J}'\hat{J})^{-1}\hat{\underline{j}}_i\,(I - \hat{\underline{j}}_i'(\hat{J}'\hat{J})^{-1}\hat{\underline{j}}_i)^{-1}\hat{\underline{j}}_i'(\hat{J}'\hat{J})^{-1}\right](-\hat{\underline{j}}_i\, e_i)$$

since $J'\underline{e} = 0$. As the 3rd factor of the 3rd term on the right hand side is a constant, i.e. $1 - h_i$,

$$\underline{\hat{\theta}}_{(i)} = \underline{\hat{\theta}} + \left[(\hat{J}'\hat{J})^{-1} + \frac{(\hat{J}'\hat{J})^{-1}\underline{\hat{j}}_i\underline{\hat{j}}_i'(\hat{J}'\hat{J})^{-1}}{(1 - h_i)}\right] (-\underline{\hat{j}}_i\, e_i)$$

where $h_i = \underline{\hat{j}}_i' (\hat{J}'\hat{J})^{-1}\underline{\hat{j}}_i$, i.e.

$$\hat{\theta}_{(i)} = \underline{\hat{\theta}} - \frac{I - I\underline{\hat{j}}_i'(\hat{J}'\hat{J})^{-1}\underline{\hat{j}}_i + (\hat{J}'\hat{J})^{-1}\underline{\hat{j}}_i\underline{\hat{j}}_i'}{1 - h_i} (\hat{J}'\hat{J})^{-1}\underline{\hat{j}}_i e_i$$

$$= \underline{\hat{\theta}} - \frac{(\hat{J}'\hat{J})^{-1}\underline{\hat{j}}_i\, e_i}{1 - h_i}$$

---

# APPENDIX D

## COMPUTER PROGRAM CODES

D.1 XAMPLE.FOR is the main program calling on SOLVE.FOR where the latter calls on EVAL.FOR. Note that another routine not given here is GAUSSJ.FOR which is a routine taken directly from NUMERICAL RECIPES (see chapter four). A further noteworthy remark is that SOLVE.FOR contains in itself several subroutines taken from NUMERICAL RECIPES, e.g. SVDCMP, SUBKSB and so forth.

The compilation will therefore consist of XAMPLE.FOR SOLVE.FOR EVAL.FOR GAUSSJ.FOR. It will be noticed that SOLVE.FOR is somewhat cluttered with so-called "derivative checks". This is necessary in order to be in full control of the calculations in EVAL.FOR. The reason is that wrong analytical derivatives are of the main causes of calculations running off track.

The code for XAMPLE.FOR SOLVE.FOR and EVAL.FOR follows. GAUSSJ.FOR can be found in the above mentioned reference.



```
C     PROGRAM XAMPLE(INPUT,OUTPUT,TAPE5=INPUT,TAPE6=OUTPUT)
C
C THE ANALYTICAL 1ST AND 2ND DERIVATIVES FOR THE MODEL MUST
C BE GIVEN IN THE SUBROUTINE EVAL. THE 1ST CALL FROM XAMPLE
C IS TO SOLVE C AND SOLVE DETERMINES THE VALUES OF THE THETAS
C CALLING ON EVAL
C
      EXTERNAL EVAL
C
      INTEGER LUNOUT,NPTS,NP,NITS,ND1,I,J,JOB,
     +IFAIL
      REAL YT(30,100),PARAM(6),VARC(5,5),DERIVS(30,6),
     +     U(6,6),W(6),V(6,6),WK1(30),DIFF(30),Y(30)
      REAL X(30,5),ALPHA,RSS,VAR,FLOAT,JAJ(6,6)
C     X VALUES ARE PASSED TO SUBROUTINE EXAL IN A COMMON BLOCK
```

```
      COMMON X
C     INITIAL PARAMETER ESTIMATES
      DATA PARAM(1)/4.0E-1/,PARAM(2)/4.0E-1/
      DATA PARAM(3)/4.5E-1/,PARAM(4)/4.5E-1/
      DATA PARAM(5)/5.0E-1/,PARAM(6)/1.5E+1/
C     VALUES FOR CONSTANTS
      DATA LUNOUT/6/,NPTS/30/,NP/6/,NITS/12/,ND1/6/
      DATA ALPHA/1.0E-3/
      DO 3 IK=1,NPTS
3     Y(IK)=0.0
C
C DATA ARE ENTERED USING AN INFILE. THE FILE ANLNOR1.DAT IS A
C 30X100FILE TO WHICH A 30X5FILE IS ANOTATED. THE LATTER
C FILE CONTINS THE X-MATRIX AND THE FIRST MATRIX CONTAINS
C 100 SAMPLES OF 30 Y-VALUES EACH SIMULATED FROM X. THE
C COMPLETE FILE IS READ, BUT ONLY THE 1ST SAMPLE USED AS AN
C EXAMPLE.
C
C NOTE THAT TWO EXTRA EXTERNAL FILES MUST BE CREATED FOR
C SOME RESULTS TO BE DUMPED INTO.
C
      OPEN(1,FILE='ANLNOR1.DAT',STATUS='OLD')
C
C DERIVS IS THE JACOBIAN, DEREXT IS THE EXTENDED FILE
C CONTAINING THE Y-VALUES AS WELL
C
      OPEN(1019,FILE='DERIVS.DAT',STATUS='OLD')
      OPEN(1020,FILE='DEREXT.DAT',STATUS='OLD')
       DO 100 I=1,NPTS
C
C THE FIRST SIX LINES IN THE DATA FILE CONTAIN SOME REMARKS
C
          IF(I .GT. 1)GO TO 14
5     READ(1,6)(YT(I,J),J=1,100),(X(I,J),J=1,5)
6     FORMAT(6(/),105F12.5)
      GO TO 18
14    READ(1,15)(YT(I,J),J=1,100),(X(I,J),J=1,5)
15    FORMAT(105F12.5)
18    Y(I)=YT(I,1)
100   CONTINUE
C DATA ENTRY COMPLETED; YT IS ENTRY DUMMY FOR Y

C      A. FIND THE LEAST SQUARES PARAMETERS ESTIMATES
          JOB=11111
        CALL SOLVE(Y,PARAM,ALPHA,NITS,NPTS,NP,ND1,LUNOUT,JOB,EVAL,
     + VARC,RSS,IFAIL,DERIVS,U,V,V,DIFF,WK1,JAJ)
      IF(IFAIL.NE.0) STOP
      VAR=RSS/FLOAT(NPTS-NP)
      WRITE(LUNOUT,35)
      WRITE(LUNOUT,40) RSS,VAR
C
C     FORMAT STATEMENTS
C
35    FORMAT(1H ,1H ,8X,3HRSS,1H ,5X,3HVAR)
```

```
40    FORMAT(1H ,F12.6,F12.8)
      DO 876 II1=1,NPTS
      WRITE(1019,1018) (DERIVS(II1,KK1),KK1=1,NP)
1018  FORMAT(6F12.6)
876   CONTINUE
      DO 978 II1=1,NPTS
      WRITE(1020,977) Y(II1),(DERIVS(II1,KK1),KK1=1,NP)
977   FORMAT(7F12.6)
978   CONTINUE
      CLOSE(1020)
      CLOSE(1019)
      CLOSE(1)
      STOP
      END


      SUBROUTINE SOLVE(Y,PARAM,ALPHA,NITS,NPTS,NP,ND1,LUNOUT,
     +JOB,EVAL,VARC,RSS,IFAIL,DERIVS,U,W,V,DIFF,WK1,JAJ)
C
C     THE ORIGINAL CODE AS GIVEN BY RATKOWSKY(1983) SUPPLIED
C     THE NECESSARY PARAMETERS IN ORDER TO DECID WHETHER ONE
C     WOULD LIKE CERTAIN IN BETWEEN RESULTS OR NOT. I OVER
C     RIDED THIS OPTION, BUT KEPT THE CODE FOR IN CASE I WOULD
C     LIKE TO MAKE USE OF IT AT A LATER STAGE
C
C NOTE THAT EXTERNAL FILES ARE USED IN CERTAIN SUBROUTINES,
C i.e. RESULTS.DAT TETAS.DAT COVM.DAT AND CORR.DAT. THEY
C MUST THEREFORE BE CREATED BEFORE THE PROGRAM IS RUN.
      EXTERNAL EVAL
      PARAMETER(NMAX=30,MMAX=6)
       INTEGER IFAIL,NPTS,NP,ND1,LUNOUT,JOB,
     +        I,J,ITER,ITASK
      REAL DERIVS(NPTS,1),VARC(ND1,NP),Y(NPTS),PARAM(NP),DIFF(NPTS),
     +WK1(NPTS),ALPHA,RSS,TEST,RESVAR,DENOM,ABS,FLOAT,
     +ZERO,ONE
      REAL YSTAR(NMAX,1),DERIT(MMAX,NMAX),DIN(MMAX,MMAX),XY(MMAX,1)
      REAL BX(MMAX),BG(MMAX),DERIV(NMAX,MMAX)
      REAL JAJ(MMAX,MMAX),JAJMI(MMAX,MMAX),RES(30),JDUM(NMAX,MMAX)
      REAL JHAT(NMAX,MMAX)
      LOGICAL CNVRGD,PRNPAR
C     THE NEXT DIMENSION NEEDED FOR SVD
       REAL U(NP,NP),W(NP),V(NP,NP)
      DATA ZERO/0.0E0/,ONE/1.0E0/
C     INITIALIZE VARIABLES
      IFAIL=0
      PRNPAR=JOB/10000.NE.0
      IF(.NOT.PRNPAR) GO TO 1111
C     PRINT A HEADING IF PARAMETER VALUES ARE REQUIRED AT EACH ITERATION
      WRITE(LUNOUT,10) (I,I=1,NP)
C     DETERMINE STARTING VALUES AND PRINT THEM
      RSS=ZERO
      DO 100 I=1,NPTS
      ITASK=0
      CALL EVAL(PARAM,I,ITASK,WK1,IFAIL)
```

```
          IF(IFAIL.NE.0) GO TO 6666
          RES(I)=Y(I)-WK1(1)
          RSS=RSS+RES(I)*RES(I)
100       CONTINUE
          WRITE(LUNOUT,20) RSS,(PARAM(I),I=1,NP)
1111      CONTINUE
           ITER=1
C
C        REPEAT LOOP TERMINATION WHEN
C        ITERATION EXCEEDS NITS,
C        OR CONVERGENCE OCCURS,
C        OR A SINGULAR MATRIX IS DETECTED
C
C
2222      CONTINUE
C         CALCULATE A DIFFERENCE VECTOR AND THE MATRIX OF FIRST
C         DERIVATIVES
          DO 200 I=1,NPTS
          ITASK=0
           CALL EVAL(PARAM,I,ITASK,WK1,IFAIL)
           IF(IFAIL.NE.0) GO TO 6666
            DIFF(I)=Y(I)-WK1(1)
           ITASK=1
           CALL EVAL(PARAM,I,ITASK,WK1,IFAIL)
           IF(IFAIL.NE.0) GO TO 6666
           DO 300 J=1,NP
           DERIVS(I,J)=WK1(J)
           DERIV(I,J)=WK1(J)
           DERIT(J,I)=WK1(J)
C
CC            WRITE(LUNOUT,199) DERIV(I,J),DERIT(J,I)
CC 199        FORMAT(1X,5X,12HDERIV(I,J)=,E12.6,3X,11HDERIT(J,I)=,E12.6)
C
300        CONTINUE
200        CONTINUE
C
C         SOLVE USING SVD AFTER CALL TO SVBKSB WK1 WILL
C         CONTAIN THE CORRECTIONS TO THE PARAMETER VALUES
C
           DO 440 I=1,NPTS
             YSTAR(I,1)=DIFF(I)
440        CONTINUE
      CALL MULT(DERIT,NP,NPTS,DERIVS,NPTS,NP,DIN,NP,NP,NP,NPTS,NP)
C
C    COPY DIN =PROD=J'J=JAJ
C
       DO 111 I=1,NP
       DO 112 J=1,NP
       JAJ(I,J)=DIN(I,J)
112    CONTINUE
111    CONTINUE

C     CHECK DERIVATIVES:-
```

```
CC       WRITE(LUNOUT,4298)
CC 4298   FORMAT(1X,18HDERIT(I,J)-MATRIKS)
CC       DO 4301 IJ1=1,NP
CC       WRITE(LUNOUT,4300) (DERIT(IJ1,KJ1),KJ1=1,NPTS)
CC 4300   FORMAT(6E12.6)
CC 4301   CONTINUE
CC       WRITE(LUNOUT,4299)
CC  4299   FORMAT(1X,18HDERIV(I,J)-MATRIKS)
CC       DO 4303 IJ2=1,NPTS
CC       WRITE(LUNOUT,4302) (DERIV(IJ2,KJ2),KJ2=1,NP)
CC 4302   FORMAT(6E8.6)
CC 4303   CONTINUE
C        END DERIVATIVE CHECK:-

       CALL MULT(DERIT,NP,NPTS,YSTAR,NPTS,1,XY,NP,1,NP,NPTS,1)

         CALL COPY(DIN,U,NP,NP,NP,NP)
          DO 430 IK=1,NP
           BX(IK)=XY(IK,1)
430      CONTINUE

         CALL GAUSSJ(DIN,NP,NP,XY,1,1)

C     COPY DIN=PROD**(-1)=JAJ**(-1)=JAJMI
         DO 113 IK=1,NP
         DO 114 JK=1,NP
         JAJMI(IK,JK)=DIN(IK,JK)
114       CONTINUE
113       CONTINUE

C
C       FURTHER CHECKS ON DERIVATIVES IF DEEMED NECESSARY
C
CC              DO 116 II2=1,NP
CC       WRITE(LUNOUT,115) (JAJMI(II2,KK2),KK2=1,NP)
CC 115    FORMAT(1X,6HJAJMI:,4X,6E8.3)
CC 116    CONTINUE
C     J[(J'J)**(-1)]J'=JDUM*J'=JHAT; CALCULATE (i)JDUM (ii)JHAT
C
      CALL MULT(DERIV,NPTS,NP,JAJMI,NP,NP,JDUM,
     +NPTS,NP,NPTS,NP,NP)
C
      CALL MULT(JDUM,NPTS,NP,DERIT,NP,NPTS,JHAT,NPTS,
     +NPTS,NPTS,NP,NPTS)
C
CC            DO 276 II3=1,NPTS
CC       WRITE(LUNOUT,275) (JHAT(II3,KK3),KK3=1,NPTS)
C
CC 275    FORMAT(1X,5HJHAT:,4X,6E8.3)
CC 276    CONTINUE
C
         CALL SVDCMP(U,NP,NP,NP,NP,W,V)
          WRITE(LUNOUT,25) (W(I),I=1,NP)
           COND=(W(1)/W(NP))**2.0
```

```
          WRITE(LUNOUT,35) COND
35        FORMAT(1H ,17HCONDITION NUMBER=,E12.6)

25        FORMAT(1H ,E12.6)
C
C          DO 483 I=1,NP
C          WRITE(LUNOUT,460) (V(I,J),J=1,NP)
C 460       FORMAT(1H ,6F12.6)
C 483       CONTINUE
C
          CALL SVBKSB(U,W,V,NP,NP,NP,NP,BX,BG)
         IFAIL=0
         DO 488 I=1,NP
         WK1(I)=BG(I)
488      CONTINUE
C
C        CORRECT PARAMETER VALUES AND
C        CHECK FOR CONVERGENCE
C
          IF(IFAIL.NE.0) GO TO 5555
          CNVRGD=.TRUE.
           DO 500 I=1,NP
            PARAM(I)=PARAM(I)+WK1(I)
             DENOM=PARAM(I)
             IF(DENOM.EQ.ZERO) DENOM=ONE
             TEST=WK1(I)/DENOM
             IF(ABS(TEST).GT.ALPHA) CNVRGD=.FALSE.
500          CONTINUE
C           DETERMINE RESIDUAL SUM OF SQUARES
            RSS=ZERO
            DO 600 I=1,NPTS
             ITASK=0
             CALL EVAL(PARAM,I,ITASK,WK1,IFAIL)
             IF(IFAIL.NE.0) GO TO 6666
             RES(I)=Y(I)-WK1(1)
             RSS=RSS+RES(I)*RES(I)
600          CONTINUE
             IF(.NOT.PRNPAR) GO TO 3333
C
C             PRINT RESIDUAL SUM OF SQUARES AND PARAMETER
C               ESTIMATES
C
              WRITE(LUNOUT,20) RSS,(PARAM(I),I=1,NP)
C
3333          CONTINUE
C            END OF REPEAT LOOP STARTING AT 2222
C              TEST FOR COMPLETION
               ITER=ITER+1
C         CALCULATE COVARIANCE MATRIX FOR PARAMETERS
          IF(NPTS.EQ.NP) GO TO 7777
           RESVAR=RSS/FLOAT(NPTS-NP)
           CALL SVDVAS(V,NP,NP,W,VARC,NP)
            DO 700 I=1,NP
            DO 800 J=1,NP
```

```
          VARC(I,J)=VARC(I,J)*RESVAR
800       CONTINUE
700       CONTINUE
C       PRINT THE RESULTS
        CALL DISPLAY(Y,VARC,PARAM,WK1,RESVAR,EVAL,NPTS,NP,ND1,
     +LUNOUT,JOB,IFAIL,JHAT)
        IF(ITER.GT.NITS) GO TO 4444
        IF(.NOT.CNVRGD) GO TO 2222
        IF(IFAIL.NE.0) GO TO 6666
        RETURN
C
4444     CONTINUE
C       NO CONVERGENCE
        WRITE(LUNOUT,30) NITS
         IFAIL=1
        RETURN
5555     CONTINUE
C       SINGULAR MATRIX
        WRITE(LUNOUT,40)
        IFAIL=2
          RETURN

6666      CONTINUE
C       USER FLAGGED ERRO IN EVAL
        WRITE(LUNOUT,50)
         IFAIL=8
         RETURN
7777     CONTINUE
C         ZERO RESIDUAL VARIANCE
          WRITE(LUNOUT,60)
          IFAIL=9
         RETURN
C
C         FORMAT STATEMENTS
10      FORMAT(//1H ,13X,19HPARAMETER ESTIMATES
     +/1H,10X,3HRSS,9(I13),/(1H ,13X,9(I13)))
20      FORMAT(1H,10(E13.6)/(1H,13X,9(E13.6)))
30      FORMAT(1H, 38H*** SUBROUTINE SOLVE  CONVERGENCE NOT,
     +15H ACHIEVED AFTER,I10,2X,10HITERATIONS//)
40      FORMAT(//1H,38H***SUBROUTINE SOLVE    SINGULAR MATRIX//)
50        FORMAT(//1H,41H*** SUBROUTINE SOLVE USER-FLAGGED ERROR,
     +19H IN SUBROUTINE EVAL//)
60      FORMAT(//1H,45H*** SUBROUTINE SOLVE   RESIDUAL VARIANCE ZERO//)
        END


       SUBROUTINE SVBKSB(U,W,V,M,N,MP,NP,B,X)
       PARAMETER (NMAX=100)
       DIMENSION U(MP,NP),W(NP),V(NP,NP),B(MP),X(NP),TMP(NMAX)
       DO 12 J=1,N
         S=0.
         IF(W(J).NE.0.)THEN
           DO 11 I=1,M
             S=S+U(I,J)*B(I)
```

```
11          CONTINUE
            S=S/W(J)
          ENDIF
          TMP(J)=S
12      CONTINUE
        DO 14 J=1,N
          S=0.
          DO 13 JJ=1,N
            S=S+V(J,JJ)*TMP(JJ)
13        CONTINUE
          X(J)=S
14      CONTINUE
        RETURN
        END
        SUBROUTINE SVDCMP(A,M,N,MP,NP,W,V)
        PARAMETER (NMAX=100)
        DIMENSION A(MP,NP),W(NP),V(NP,NP),RV1(NMAX)
        G=0.0
        SCALE=0.0
        ANORM=0.0
        DO 25 I=1,N
          L=I+1
          RV1(I)=SCALE*G
          G=0.0
          S=0.0
          SCALE=0.0
          IF (I.LE.M) THEN
            DO 11 K=I,M
              SCALE=SCALE+ABS(A(K,I))
11          CONTINUE
            IF (SCALE.NE.0.0) THEN
              DO 12 K=I,M
                A(K,I)=A(K,I)/SCALE
                S=S+A(K,I)*A(K,I)
12            CONTINUE
              F=A(I,I)
              G=-SIGN(SQRT(S),F)
              H=F*G-S
              A(I,I)=F-G
              IF (I.NE.N) THEN
                DO 15 J=L,N
                  S=0.0
                  DO 13 K=I,M
                    S=S+A(K,I)*A(K,J)
13                CONTINUE
                  F=S/H
                  DO 14 K=I,M
                    A(K,J)=A(K,J)+F*A(K,I)
14                CONTINUE
15              CONTINUE
              ENDIF
              DO 16 K= I,M
                A(K,I)=SCALE*A(K,I)
16            CONTINUE
```

```
          ENDIF
        ENDIF
        V(I)=SCALE *G
        G=0.0
        S=0.0
        SCALE=0.0
        IF ((I.LE.M).AND.(I.NE.N)) THEN
          DO 17 K=L,N
            SCALE=SCALE+ABS(A(I,K))
17        CONTINUE
          IF (SCALE.NE.0.0) THEN
            DO 18 K=L,N
              A(I,K)=A(I,K)/SCALE
              S=S+A(I,K)*A(I,K)
18          CONTINUE
            F=A(I,L)
            G=-SIGN(SQRT(S),F)
            H=F*G-S
            A(I,L)=F-G
            DO 19 K=L,N
              RV1(K)=A(I,K)/H
19          CONTINUE
            IF (I.NE.M) THEN
              DO 23 J=L,M
                S=0.0
                DO 21 K=L,N
                  S=S+A(J,K)*A(I,K)
21              CONTINUE
                DO 22 K=L,N
                  A(J,K)=A(J,K)+S*RV1(K)
22              CONTINUE
23            CONTINUE
            ENDIF
            DO 24 K=L,N
              A(I,K)=SCALE*A(I,K)
24          CONTINUE
          ENDIF
        ENDIF
        ANORM=MAX(ANORM,(ABS(W(I))+ABS(RV1(I))))
25    CONTINUE
      DO 32 I=N,1,-1
        IF (I.LT.N) THEN
          IF (G.NE.0.0) THEN
            DO 26 J=L,N
              V(J,I)=(A(I,J)/A(I,L))/G
26          CONTINUE
            DO 29 J=L,N
              S=0.0
              DO 27 K=L,N
                S=S+A(I,K)*V(K,J)
27            CONTINUE
              DO 28 K=L,N
                V(K,J)=V(K,J)+S*V(K,I)
28            CONTINUE
```

```
29          CONTINUE
          ENDIF
          DO 31 J=L,N
            V(I,J)=0.0
            V(J,I)=0.0
31        CONTINUE
        ENDIF
        V(I,I)=1.0
        G=RV1(I)
        L=I
32    CONTINUE
      DO 39 I=N,1,-1
        L=I+1
        G=W(I)
        IF (I.LT.N) THEN
          DO 33 J=L,N
            A(I,J)=0.0
33        CONTINUE
        ENDIF
        IF (G.NE.0.0) THEN
          G=1.0/G
          IF (I.NE.N) THEN
            DO 36 J=L,N
              S=0.0
              DO 34 K=L,M
                S=S+A(K,I)*A(K,J)
34            CONTINUE
              F=(S/A(I,I))*G
              DO 35 K=I,M
                A(K,J)=A(K,J)+F*A(K,I)
35            CONTINUE
36          CONTINUE
          ENDIF
          DO 37 J=I,M
            A(J,I)=A(J,I)*G
37        CONTINUE
        ELSE
          DO 38 J= I,M
            A(J,I)=0.0
38        CONTINUE
        ENDIF
        A(I,I)=A(I,I)+1.0
39    CONTINUE
      DO 49 K=N,1,-1
        DO 48 ITS=1,30
          DO 41 L=K,1,-1
            NM=L-1
            IF ((ABS(RV1(L))+ANORM).EQ.ANORM)  GO TO 2
            IF ((ABS(W(NM))+ANORM).EQ.ANORM)  GO TO 1
41        CONTINUE
1         C=0.0
          S=1.0
          DO 43 I=L,K
            F=S*RV1(I)
```

```
            IF ((ABS(F)+ANORM).NE.ANORM) THEN
              G=W(I)
              H=SQRT(F*F+G*G)
              W(I)=H
              H=1.0/H
              C= (G*H)
              S=-(F*H)
              DO 42 J=1,M
                Y=A(J,NM)
                Z=A(J,I)
                A(J,NM)=(Y*C)+(Z*S)
                A(J,I)=-(Y*S)+(Z*C)
42            CONTINUE
            ENDIF
43        CONTINUE
2         Z=W(K)
          IF (L.EQ.K) THEN
            IF (Z.LT.0.0) THEN
              W(K)=-Z
              DO 44 J=1,N
                V(J,K)=-V(J,K)
44            CONTINUE
            ENDIF
            GO TO 3
          ENDIF
          IF (ITS.EQ.30) PAUSE 'No convergence in 30 iterations'
          X=W(L)
          NM=K-1
          Y=W(NM)
          G=RV1(NM)
          H=RV1(K)
          F=((Y-Z)*(Y+Z)+(G-H)*(G+H))/(2.0*H*Y)
          G=SQRT(F*F+1.0)
          F=((X-Z)*(X+Z)+H*((Y/(F+SIGN(G,F)))-H))/X
          C=1.0
          S=1.0
          DO 47 J=L,NM
            I=J+1
            G=RV1(I)
            Y=W(I)
            H=S*G
            G=C*G
            Z=SQRT(F*F+H*H)
            RV1(J)=Z
            C=F/Z
            S=H/Z
            F= (X*C)+(G*S)
            G=-(X*S)+(G*C)
            H=Y*S
            Y=Y*C
            DO 45 NM=1,N
              X=V(NM,J)
              Z=V(NM,I)
              V(NM,J)= (X*C)+(Z*S)
```

```
              V(NM,I)=-(X*S)+(Z*C)
45            CONTINUE
              Z=SQRT(F*F+H*H)
              W(J)=Z
              IF (Z.NE.0.0) THEN
                Z=1.0/Z
                C=F*Z
                S=H*Z
              ENDIF
              F= (C*G)+(S*Y)
              X=-(S*G)+(C*Y)
              DO 46 NM=1,M
                Y=A(NM,J)
                Z=A(NM,I)
                A(NM,J)= (Y*C)+(Z*S)
                A(NM,I)=-(Y*S)+(Z*C)
46            CONTINUE
47          CONTINUE
            RV1(L)=0.0
            RV1(K)=F
            W(K)=X
48        CONTINUE
3         CONTINUE
49      CONTINUE
        RETURN
        END
        SUBROUTINE SVDVAS(V,MA,NP,W,CVM,NCVM)
C       SVDVAS RETURNS COVARIANCE MATRIX - X'X IS USED

        PARAMETER (MMAX=20)
        DIMENSION V(NP,NP),W(NP),CVM(NCVM,NCVM),WTI(MMAX)
        DO 11 I=1,MA
          WTI(I)=0.
C IF X IS NON-SINGULAR MULTIPLY BY 1./(W(I)
          IF(W(I).NE.0.) WTI(I)=1./W(I)
11      CONTINUE
        DO 14 I=1,MA
          DO 13 J=1,I
            SUM=0.
            DO 12 K=1,MA
              SUM=SUM+V(I,K)*V(J,K)*WTI(K)
12          CONTINUE
            CVM(I,J)=SUM
            CVM(J,I)=SUM
13        CONTINUE
14      CONTINUE
        RETURN
        END
        SUBROUTINE DISPLAY(Y,VARC,PARAM,WK1,VAR,EVAL,NPTS,NP,ND1,
     +                     LUNOUT,JOB,IFAIL,JHAT)
C
C       EXTERNAL EVAL
         INTEGER JOB,MOD,ND1,NP,I,J,NPTS,IFAIL,ITASK,LUNOUT
         REAL Y(1),VARC(ND1,1),PARAM(1),WK1(1),VAR,SQRT,T,SE,
```

```
     +      YHAT,RES(30),DI(30),DFFITS(30),JHAT(30,30)
       LOGICAL PRNTAB,PRNVCV,PRNCOR,PRNFIT
C
      OPEN(51,FILE='RESULTS.DAT',STATUS='OLD')
      OPEN(52,FILE='TETAS.DAT',STATUS='OLD')
      OPEN(53,FILE='COVM.DAT',STATUS='OLD')
      OPEN(54,FILE='CORR.DAT',STATUS='OLD')
C
        PRNTAB=MOD(JOB,10000)/1000.NE.0
        PRNVCV=MOD(JOB,1000)/100.NE.0
        PRNCOR=MOD(JOB,100)/10.NE.0
        PRNFIT=MOD(JOB,10).NE.0
C
        IF(.NOT.PRNTAB) GO TO  1111
C       PRINT A TABLE OF ESTIMATES
        WRITE(LUNOUT,10)
        DO 100 I=1,NP
        SE=SQRT(VARC(I,I))
        T=PARAM(I)/SE
        WRITE(LUNOUT,20) I,PARAM(I),SE,T
        WRITE(52,20) I,PARAM(I),SE,T
100     CONTINUE
C
1111     CONTINUE

        IF(.NOT.PRNVCV) GO TO 2222
C        PRINT COVARIANCE MATRIX
         WRITE(LUNOUT,30)
         DO 200 I=1,NP
         WRITE(LUNOUT,40) I,(VARC(I,J),J=1,I)
         WRITE(53,40) I,(VARC(I,J),J=1,I)
200      CONTINUE
2222     CONTINUE
          IF(.NOT.PRNCOR) GO TO 3333
          WRITE(LUNOUT,50)
C      PRINT CORRELATION MATRIX
        DO 300 I=1,NP
         DO 400 J=1,I
         WK1(J)=VARC(I,J)/SQRT(VARC(I,I)*VARC(J,J))
400      CONTINUE
         WRITE(LUNOUT,60) I,(WK1(J),J=1,I)
         WRITE(54,60) I,(WK1(J),J=1,I)
300       CONTINUE
C
3333     CONTINUE
          IF(.NOT.PRNFIT) GO TO 4444
           WRITE(LUNOUT,70)
C
C  DI=COOK'S MEASURE APPR.; DFFITS=DFITS**2
C
            CURES2=0
            ZZ=0
            DO 500 I=1,NPTS
            ITASK=0
```

```
            CALL EVAL(PARAM,I,ITASK,WK1,IFAIL)
            IF(IFAIL.NE.0) RETURN
             YHAT=WK1(1)
             RES(I)=Y(I)-YHAT
             DI(I)=(JHAT(I,I)/(1.-JHAT(I,I)))*RES(I)
             DFFITS(I)=DI(I)*RES(I)/(VAR*(1.-JHAT(I,I)))
             CU=RES(I)*RES(I)
             CURES2=CU+ZZ
             ZZ=CURES2
             WRITE(LUNOUT,80) I,Y(I),YHAT,RES(I),
     +JHAT(I,I),DFFITS(I)
             WRITE(51,80) I,Y(I),YHAT,RES(I),
     +JHAT(I,I),DFFITS(I)
500         CONTINUE
4444        CONTINUE
C           FORMAT STATEMENTS
10        FORMAT(////1H ,10X,34HPARAMETER ESTIMATES AT CONVERGENCE/
     +/1H ,9HPARAMETER,4X,8HESTIMATE,11X,2HSE,11X,1HT)
20          FORMAT(1H, I5,4X,E12.6,1X,E12.6,2X,F10.2)

30        FORMAT(////1H ,10X,36HPARAMETER VARIANCE-COVARINCE MATRIX//)
40        FORMAT(1H ,I5,8(2X,E12.6)/(1H, 5X,8(2X,F12.6)))
50        FORMAT(////1H ,10X,22HPARAMETER CORRELATIONS//)
60        FORMAT(1H ,I5,8(2X,F12.6)/(1H ,5X,8(2X,F12.6)))
70        FORMAT(////1H ,1X,4HUNIT,6X,1HY,7X,6HFITTED,6X,
     + 8HRESIDUAL,5X,9HJHAT(I,I),2X,9HDFFITS**2)
80        FORMAT(1H ,I4,7E12.6)
          CLOSE(51)
          END
          SUBROUTINE MULT(MAT1,M1,N1,MAT2,M2,N2,PROD,M3,N3,II,KK,JJ)
          REAL MAT1(M1,N1),MAT2(M2,N2),PROD(M3,N3)
           DO 7000 I=1,II,1
             DO 705 J=1,JJ,1
               PROD(I,J)=0.0
               DO 7010 K=1,KK,1
                   PROD(I,J)=PROD(I,J)+MAT1(I,K)*MAT2(K,J)
7010           CONTINUE
705          CONTINUE
7000       CONTINUE
           RETURN
           END
          SUBROUTINE TRANP(A,B,NP,MP,N,M)
          REAL A(NP,MP),B(MP,NP)
          DO 8000 I=1,N
          DO 8010 J=1,M
          B(J,I)=A(I,J)
8010      CONTINUE
8000      CONTINUE
          RETURN
          END
          SUBROUTINE COPY(A,B,NP,MP,IA,JB)
C         COPIES A(IA,JB) INTO B AND LOSES PREVIOUS B
           REAL A(NP,MP)
           REAL B(NP,MP)
```

```
      DO 10 I=1,IA
      DO 20 J=1,JB
      B(I,J)=A(I,J)
20    CONTINUE
10    CONTINUE
      RETURN
      END


      SUBROUTINE EVAL(PARAM,I,ITASK,WK1,IFAIL)
      INTEGER I,ITASK,IFAIL
      REAL PARAM(6),WK1(21),X
      REAL XX1,XX2,XX3,XX4,XX5,TT1,TT2
      COMMON X(30,5)
      IFAIL=0
      IF(ITASK.NE.0) GO TO 1111
C     PREDICTED VALUE IS COMPUTED
      XX1=X(I,1)**PARAM(1)
      XX2=X(I,2)**PARAM(2)
      XX3=X(I,3)**PARAM(3)
      XX4=X(I,4)**PARAM(4)
      XX5=X(I,5)**PARAM(5)
      TT1=XX1*XX2*XX3*XX4*XX5*PARAM(6)
      WK1(1)=TT1
      RETURN
1111  CONTINUE
      IF(ITASK.NE.1) GO TO 2222
C     FIRST DERIVATIVES
      XX1=X(I,1)**PARAM(1)
      XX2=X(I,2)**PARAM(2)
      XX3=X(I,3)**PARAM(3)
      XX4=X(I,4)**PARAM(4)
      XX5=X(I,5)**PARAM(5)
      TT1=XX1*XX2*XX3*XX4*XX5*PARAM(6)
      WK1(1)=TT1*LOG(X(I,1))
      WK1(2)=TT1*LOG(X(I,2))
      WK1(3)=TT1*LOG(X(I,3))
      WK1(4)=TT1*LOG(X(I,4))
      WK1(5)=TT1*LOG(X(I,5))
      WK1(6)=TT1/PARAM(6)
      RETURN
2222  CONTINUE
      IF(ITASK.NE.2) GO TO 3333
C      SECOND DERIVATIVES
      XX1=X(I,1)**PARAM(1)
      XX2=X(I,2)**PARAM(2)
      XX3=X(I,3)**PARAM(3)
      XX4=X(I,4)**PARAM(4)
      XX5=X(I,5)**PARAM(5)
      TT2=XX1*XX2*XX3*XX4*XX5*PARAM(6)
       WK1(1)=(LOG(X(I,1))**2.)*TT2
       WK1(2)=LOG(X(I,2))*LOG(X(I,1))*TT2
       WK1(3)=LOG(X(I,3))*LOG(X(I,1))*TT2
       WK1(4)=LOG(X(I,4))*LOG(X(I,1))*TT2
```

```
      WK1(5)=LOG(X(I,5))*LOG(X(I,1))*TT2
      WK1(6)=LOG(X(I,1))*TT2/PARAM(6)
      WK1(7)=(LOG(X(I,2))**2.)*TT2
      WK1(8)=LOG(X(I,3))*LOG(X(I,2))*TT2
      WK1(9)=LOG(X(I,4))*LOG(X(I,2))*TT2
      WK1(10)=LOG(X(I,5))*LOG(X(I,2))*TT2
      WK1(11)=LOG(X(I,2))*TT2/PARAM(6)
      WK1(12)=(LOG(X(I,3))**2.)*TT2
      WK1(13)=LOG(X(I,4))*LOG(X(I,3))*TT2
      WK1(14)=LOG(X(I,5))*LOG(X(I,3))*TT2
      WK1(15)=LOG(X(I,3))*TT2/PARAM(6)
      WK1(16)=(LOG(X(I,4))**2.)*TT2
      WK1(17)=LOG(X(I,5))*LOG(X(I,4))*TT2
      WK1(18)=LOG(X(I,4))*TT2/PARAM(6)
      WK1(19)=(LOG(X(I,5))**2.)*TT2
      WK1(20)=LOG(X(I,5))*TT2/PARAM(6)
      WK1(21)=0
      RETURN
3333   CONTINUE
C     ERROR SET IFAIL TO NONZERO VALUE
       IFAIL=1
       RETURN
       END
```

D.2 RFGPRN.FOR analyses the data set or in this thesis, the jacobian into a singular value decomposition. The subroutines in this case are all from NUMERICAL RECIPES, so that only the main program code is supplied.

The compilation consists of RFGPRN.FOR SVDCMP.FOR where SVDCMP.FOR is now an external subroutine called from RFGPRN.FOR. Again SVDCMP.FOR is a NUMERICAL RECIPE routine.

```
      PARAMETER(NP=30,MP=6)
      CHARACTER*12 FNAME
      DIMENSION X(NP,MP),X1(NP,MP),A(NP,MP)
      DIMENSION U(NP,MP),W(MP),V(MP,MP)
      DIMENSION AP(MP,NP),S(MP,MP),AVE(MP),SD(MP)
      DIMENSION C1(NP),B1(MP),X2(MP,MP)
      DIMENSION X1T(MP,NP),X1IN(MP,MP)
      DIMENSION XM(MP,MP)
      DIMENSION F(NP,MP),G(MP,MP),F1(NP,MP),G1(MP,MP)
      DIMENSION FRSUM(NP),FCSUM(MP),GRSUM(MP),GCSUM(MP)
      DIMENSION F2(NP,MP),G2(MP,MP)
      DIMENSION TOT(MP),PHI(MP,MP)
C
      REAL FRTOT,FCTOT
```

```
      DIMENSION FRRAT(NP),FCRAT(MP)
C
      INTEGER NE,NS
       WRITE(*,*) 'ENTER FILENAME AS CASDAT 6 CHARACTERS'
      READ(*,80) FNAME
80    FORMAT(A12)
      WRITE(*,70) FNAME
70    FORMAT(1X,A12)
      OPEN(15,FILE=FNAME,STATUS='OLD')
      OPEN(16,FILE='SVD.DAT',STATUS='OLD')
      OPEN(17,FILE='LES.DAT',STATUS='OLD')
      OPEN(18,FILE='RES.DAT',STATUS='OLD')
      OPEN(19,FILE='INF.DAT',STATUS='OLD')
      OPEN(20,FILE='F1F2.DAT',STATUS='OLD')
      OPEN(21,FILE='G1G2.DAT',STATUS='OLD')
      WRITE(*,*) 'ENTER N=NO. OF ROWS AND M=NO. OF COLUMNS,'
      WRITE(*,*)'RC AND IP, NS=SAMPLE NO.'
      READ(*,90) N,M,RC,IP,NS
90    FORMAT(2I5,F5.1,2I5)
      WRITE(*,95) N,M,RC,IP,NS
95     FORMAT(1X,2I5,F5.2,2I5)
      READ(15,*) ((X(K,L),L=1,M),K=1,N)

      WRITE(*,151) ((X(K,L),L=1,M),K=1,N)
151   FORMAT(1X,6(1XF12.6))

      DO 110 K=1,N
       DO 120 L=1,M
       X1(K,L)=X(K,L)
120    CONTINUE
110    CONTINUE

      MQ=M

      CALL TRANP(X1,X1T,NP,MP,N,M)
      CALL MULT(X1T,MP,NP,X1,NP,MP,X1IN,MP,MP,M,N,M)

      CALL COPY(X,A,NP,MP,N,M)
C     XM=X1IN=X'X, X1IN REPLACE BY X'X INVERSE
      DO 61 I=1,M
      DO 63 J=1,M
      XM(I,J)=X1IN(I,J)
63    CONTINUE
61    CONTINUE


55    CONTINUE


       CALL CNTRAL(A,NP,MP,N,M,AVE,M)
      CALL TRANP(A,AP,NP,MP,N,M)
      CALL MULT(AP,MP,NP,A,NP,MP,S,MP,MP,M,N,M)
        WRITE(*,*) '      MEANS    '
       WRITE(*,135) (AVE(K),K=1,M)
```

```
135     FORMAT(1X,6F12.6)

       DO 140 K=1,M
        SD(K)=SQRT(S(K,K)/FLOAT(N-1))
140    CONTINUE
           WRITE(*,*) '   STANDARD DEVIATIONS  '
           WRITE(*,135) (SD(K),K=1,M)
       CALL COPY(A,U,NP,MP,N,M)
       CALL SVDCMP(U,N,M,NP,MP,W,V)
       WRITE(*,*)      '       EIGEN VALUES OF CENTERED MATRIX '
       WRITE(*,360) (W(I),I=1,M)
360    FORMAT(1X,6E12.6)
c       CALL CORR(A,S,C,NP,MP,N,M)
c       CALL COPY(C,U,NP,MP,N,M)
c       CALL SVDCMP(U,N,M,NP,MP,W,V)
C        DO 4444 I=1,M
C        C1(I)=0.0
C  4444       B1(I)=0.0
C       CALL SVDSORT(U,W,V,N,M,NP,MP,C1,B1)
C       WRITE(*,*)
C       WRITE(*,*) 'SORTED EIGEN VALUES OF STANDARDIZED MATRIX '
C       WRITE(*,*)
C       WRITE(*,360) (W(I),I=1,M)
        DO 133 J=1,M

          DO 144 I=1,M
       WRITE(*,1441) V(I,J),W(J)
1441   FORMAT(1X,7HV(I,J)=,E12.6,3X,5HW(J)=,E12.6)
          G(I,J)=V(I,J)*W(J)
144       CONTINUE
133       CONTINUE
        CALL APPROX(G,M,MP,IP,PHI,TOT)
        WRITE(*,*)'J=U V VTRANSP ;U,V,V FOLLOW, NOT VTRANSP'
         DO 5 I=1,N
          write(*,6) i,(u(i,j),j=1,mq)
5         WRITE(16,6) I,(U(I,J),J=1,MQ)
          write(*,*)
          WRITE(16,*)
          write(*,7) (w(i),i=1,mq)
          WRITE(16,7) (W(I),I=1,MQ)
          write(*,*)
7          FORMAT(8F9.4)
          DO 8 I=1,MQ
           write(*,6) i,(v(i,j),j=1,mq)
8          WRITE(16,6) I,(V(I,J),J=1,MQ)
6       FORMAT(I3,8F12.6)
         DO 11 J=1,MQ
         DO 12 I=1,N
         F(I,J)=U(I,J)*W(J)
12      CONTINUE
11      CONTINUE
         DO 32 I=1,N
32       FRSUM(I)=0.0
         DO 33 I=1,MQ
```

```
      FCSUM(I)=0.0
      GRSUM(I)=0.0
33    GCSUM(I)=0.0
      DO 34 I=1,N
      DO 35 J=1,MQ
       FRSUM(I)=FRSUM(I) +F(I,J)**2
35    CONTINUE
34    CONTINUE

      DO 36 J=1,MQ
      DO 37 I=1,N
       FCSUM(J)=FCSUM(J)+F(I,J)**2
37    CONTINUE
       WRITE(*,69) FCSUM(J)
69      FORMAT(3X,F20.10)
36    CONTINUE
          DO 42 I=1,N
          DO 43 J=1,MQ
           F1(I,J)=F(I,J)**2/FRSUM(I)
43      CONTINUE
42      CONTINUE

         DO 44  J=1,MQ
         DO 45   I=1,N
         F2(I,J)=F(I,J)**2/FCSUM(J)
45      CONTINUE
44       CONTINUE


     DO 13 J=1,MQ

       DO 14 I=1,MQ
       G(I,J)=V(I,J)*V(J)
14    CONTINUE
13    CONTINUE
       DO 557 I=1,MQ
       DO 556 J=1,MQ
       GRSUM(I)=GRSUM(I)+G(I,J)**2
556    CONTINUE
557     CONTINUE
       DO 577 J=1,MQ
       DO 587 I=1,MQ
        GCSUM(J)=GCSUM(J)+G(I,J)**2

587      CONTINUE
577      CONTINUE
       DO 46 I=1,MQ
       DO 47 J=1,MQ
       G1(I,J)=G(I,J)*G(I,J)/GRSUM(I)
47     CONTINUE
46     CONTINUE
       DO 48 J=1,MQ
       DO 49 I=1,MQ
       G2(I,J)=G(I,J)*G(I,J)/GCSUM(J)
```

```
49          CONTINUE
48           CONTINUE
C
            FCTOT=0.0
            FRTOT=0.0
            DO 1150 I=1,N
            FRTOT=FRTOT+FRSUM(I)
1150        CONTINUE
            DO 1250 J=1,M
            FCTOT=FCTOT+FCSUM(J)
1250        CONTINUE
            DO 1151 I=1,N
            FRRAT(I)=FRSUM(I)/FRTOT
1151        CONTINUE
            DO 1251 J=1,M
            FCRAT(J)=FCSUM(J)/FCTOT
1251        CONTINUE
C
            write(*,*)'F1=F**2/FRSUM,FRSUM,FRSUM/FRTOT'
             write(*,*)'F2=F**2/FCSUM,FCSUM,FCSUM/FCTOT'

             DO 50 I=1,N
             write(*,66) I,(F1(I,J),J=1,MQ),FRSUM(I),FRRAT(I)
50           WRITE(20,66) I,(F1(I,J),J=1,MQ),FRSUM(I),FRRAT(I)
66           FORMAT(I3,1x,6F7.3,1X,F15.6,1x,F6.4)
             write(*,*)
             DO 51 I=1,N
             write(*,6) I,(F2(I,J),J=1,MQ)
51           WRITE(20,6) I,(F2(I,J),J=1,MQ)
             write(*,*)
             WRITE(20,68) (FCSUM(J),J=1,MQ)
             WRITE(*,68) (FCSUM(J),J=1,MQ)
             write(*,*)
             WRITE(20,68) (FCRAT(J),J=1,MQ)
             WRITE(*,68) (FCRAT(J),J=1,MQ)
68           FORMAT(3X,4F20.10,/,3X,4F20.10)
             write(*,*)'G1 AND G2 FOLLOW'
             DO 52 I=1,MQ
             write(*,6) I,(G1(I,J),J=1,MQ)
52           WRITE(21,6) I,(G1(I,J),J=1,MQ)
             write(*,*)
              DO 53 I=1,MQ
             write(*,6) I,(G2(I,J),J=1,MQ)
53           WRITE(21,6) I,(G2(I,J),J=1,MQ)
             write(*,*)'F(I,J) AND G(I,J) FOLLOW'
             DO 15 I=1,N
             write(*,6) I,(F(I,J),J=1,MQ)
15            WRITE(17,6) I,(F(I,J),J=1,MQ)
            DO 16 I=1,MQ
            write(*,6) I,(G(I,J),J=1,MQ)
16          WRITE(17,6) I,(G(I,J),J=1,MQ)
C
C
          DO 4445 I=1,M
```

```
        C1(I)=0.0
4445       B1(I)=0.0
        CALL SVDSORT(U,W,V,N,M,NP,MP,C1,B1)
       WRITE(*,*)
       WRITE(*,*) 'SORTED EIGEN VALUES OF CENTERED MATRIX '
       WRITE(*,*)
       WRITE(*,360) (W(I),I=1,M)


         WRITE(*,*) 'EIGEN VECTORS'
         DO 630 I=1,MQ
         WRITE(*,360) (U(I,J),J=1,MQ)
630      CONTINUE
        COND=W(1)/W(MQ)
        WRITE (*,*) 'CONDITION NUMBER '
        WRITE(*,380) COND
380     FORMAT(1X,F15.6)


       DO 700 L=MQ,1,-1
C       ELIMINATE EIGEN VALUES UNTIL CONDITION NUMBER <30
       CRIT=W(1)/W(L)
       IF (CRIT-30) 710,710,700
700    CONTINUE
710    IF (L-MQ) 720,725,725

725    WRITE (*,*) 'NO EIGEN VALUES ELIMINATED'
       GO TO 510
720    DO 730 I=L+1,MQ
       W(I)=0.0
730    CONTINUE


         DO 740 I=1,MQ
         DO 750 J=1,MQ
         DO 760 K=1,L
         X2(I,J)=X2(I,J)+V(I,K)*V(J,K)/W(K)
760      CONTINUE
750      CONTINUE
740      CONTINUE


         WRITE(*,*) 'NUMBER OF EIGEN VALUES ELIMINATED '
         NE=MQ-L
         WRITE(*,800) NE
800      FORMAT(1X,I5,/)

6660     CONTINUE
510        CLOSE(5)
        STOP
        END
       SUBROUTINE CORR(A,S,C,NP,MP,N,M)
C      A IS CENTERED MATRIX S IS WISHART MATRIC C IS RETURNED AS STAND. MAT.
       REAL A(NP,MP),S(MP,MP),C(NP,MP)
```

```
      DO 100 I=1,N
      DO 110 J=1,M
      C(I,J)=0.0
110    CONTINUE
100   CONTINUE
      DO 16 J=1,M
      DO 17 I=1,N
      C(I,J)=(A(I,J)/SQRT(S(J,J)))
17    CONTINUE
16    CONTINUE
      RETURN
      END


      SUBROUTINE MULT(MAT1,M1,N1,MAT2,M2,N2,PROD,M3,N3,II,KK,JJ)
      REAL MAT1(M1,N1),MAT2(M2,N2),PROD(M3,N3)
       DO 7000 I=1,II,1
         DO 705 J=1,JJ,1
           PROD(I,J)=0.0
           DO 7010 K=1,KK,1
               PROD(I,J)=PROD(I,J)+MAT1(I,K)*MAT2(K,J)
7010       CONTINUE
705      CONTINUE
7000   CONTINUE
       RETURN
       END
      SUBROUTINE CNTRAL(MAT,NP,MP,DIM1,DIM2,AVE,N1)
      REAL MAT(NP,MP)
      INTEGER DIM1 ,DIM2
      REAL AVE(N1)
      DO 6000 J=1,DIM2,1
      AVE(J)=0.0
      DO 605 I=1,DIM1,1
       AVE(J)=AVE(J)+MAT(I,J)
605    CONTINUE
       AVE(J)=AVE(J)/FLOAT(DIM1)
6000  CONTINUE
C      AVE(J) NOW CONTAINS AVERAGE OF ELEMENTS IN EACH COLUMN
      DO 6010 I=1,DIM1,1
        DO 6022 J=1,DIM2,1
          MAT(I,J)=MAT(I,J)-AVE(J)
6022    CONTINUE
6010    CONTINUE
        RETURN
        END
      SUBROUTINE TRANP(A,B,NP,MP,N,M)
      REAL A(NP,MP),B(MP,NP)
      DO 8000 I=1,N
      DO 8010 J=1,M
      B(J,I)=A(I,J)
8010  CONTINUE
8000  CONTINUE
      RETURN
      END
```

```
      SUBROUTINE SUM(A,B,C,NP,N)
C     PUTS SUMS OF SQARE MATRICES A+B INTO C,LOOSING OLD C
      REAL  A(NP,NP)
      REAL B(NP,NP)
      REAL C(NP,NP)
      DO 10 I=1,N
        DO 20 J=1,N
         C(I,J)=A(I,J)+B(I,J)
20    CONTINUE
10    CONTINUE
      RETURN
      END
      SUBROUTINE COPY(A,B,NP,MP,IA,JB)
C     COPIES A(IA,JB) INTO B AND LOSES PREVIOUS B
       REAL A(NP,MP)
       REAL B(NP,MP)
      DO 10 I=1,IA
      DO 20 J=1,JB
      B(I,J)=A(I,J)
20    CONTINUE
10    CONTINUE
      RETURN
      END
      SUBROUTINE SVDSORT(U,W,V,N,M,NP,MP,C,B)
      DIMENSION U(NP,MP),W(MP),V(MP,MP)
      DIMENSION C(NP),B(MP)
      DO 90 K=1,M-1
      DO 100 J=1,M-K
      IF(W(J).LT.W(J+1)) THEN
       HOLD=W(J)
       DO 110 L=1,N
       C(L)=U(L,J)
110    CONTINUE
       DO 120 L=1,M
        B(L)=V(L,J)
120     CONTINUE
       W(J)=W(J+1)
       DO  130 L=1,N
       U(L,J)=U(L,J+1)
130    CONTINUE
       DO 140 L=1,M
       V(L,J)=V(L,J+1)
140    CONTINUE
       W(J+1)= HOLD
       DO 150 L=1,N
       U(L,J+1)=C(L)
150    CONTINUE
       DO 160 L=1,M
       V(L,J+1)=B(L)
160    CONTINUE
       END IF
100      CONTINUE
90       CONTINUE
        RETURN
```

```
      END



      SUBROUTINE BOXCOX(X,IY,NNS,MMS,BC,OFFSET)
C- - - - - - - - - - - - - - - - - - - - - - -30-3-85
C   BOX-COX EXTENDED POWER FAMILY OF TRANSFORMATIONS OF DEPENDENT
C   VARIABLE -  SEE COOK & WEISBERG 1982 p 60
C- - - - - - - - - - - - - - - - - - - - - - -LGU
      REAL X(NNS,MMS)
      OFF=0.
      IF(OFFSET.GT.-9998.)OFF=OFFSET
      WRITE(6,101) IY,BC,OFF
101   FORMAT(' VARIABLE',I4,' TRANSFORMED BY BOX-COX EXTENDED POWER'
     &,' FAMILY WITH PARAMETERS POWER =',F5.2,', OFFSET =',F5.2)
      IF(BC.NE.0) THEN
      DO 1 I=1,NNS
      DIOFF=X(I,IY)+OFF
      IF(DIOFF .LT.0.0)WRITE(6,100)DIOFF




1     X(I,IY)=(DIOFF**BC-1.)/BC
100   FORMAT(' ERROR CAUSED BY NEGATIVE VALUE OF DEPENDENT VARIBLE'
     &,F10.6)
      ELSE
      DO 2 I=1,NNS
      DIOFF=X(I,IY)+OFF
      IF(DIOFF.LE.0.0)WRITE(6,100)DIOFF
2     X(I,IY)=ALOG(DIOFF)
      ENDIF
      RETURN
      END
C--4----------------------------------------------------------------
      SUBROUTINE MAPRNT(X,IR,IC,N,M)
C- - - - - - - - - - - - - - - - - - - - - - -18-3-85
CPRINTS AN  NxM  MATRIX WITHOUT ANNOTATION
C- - - - - - - - - - - - - - - - - - - - - - -LGU
      REAL X(IR,IC)
      M9=(M+8)/9
      DO 1 K=1,M9
      I1=(K-1)*9+1
      I2=MIN(K*9,M)
      WRITE(6,51) (I,I=I1,I2)
      DO 1 J=1,N
      WRITE(6,50)J,(X(J,I),I=I1,I2)
1     CONTINUE
51    FORMAT('0',17X,9I7)
50    FORMAT(' ',13X,I4,4X,9F7.4)
```

```
52    format(' ',10f8.4)
      do 2 j=1,n
2     write(15,52) (x(j,i),i=1,m)
      RETURN
      END
c--45-----------------------------------------------------------------
      subroutine approx(g,m,mp,ip,phi,tot)
c- - - - - - - - - - - - - - - - - - - - - - - - - -07-07-89
c           computes cosines of angles approximating correlations
c           phi(i,j)= f^t f / norm f norm f
c- - - - - - - - - - - - - - - - - - - - - - - - - -LGU
      real g(mp,mp),phi(mp,mp),tot(mp)
      do 2 j=1,m
      tot(j)=0.
      do 1 k=1,ip
1     tot(j)=tot(j)+g(j,k)*g(j,k)
2     tot(j)=sqrt(tot(j))
      WRITE(*,5) TOT(J)
5     FORMAT(1X,7HTOT(J)=,E12.6)
      do 4 j=1,m
      do 4 i=j,m
      phi(i,j)=0.
      do 3 k=1,ip
3     phi(i,j)=phi(i,j) + g(i,k)*g(j,k)
      phi(i,j)=phi(i,j)/(tot(i)*tot(j))
4     phi(j,i)=phi(i,j)
      write(6,100) (tot(j), j=1,m)
100   format(' Lengths of column vectors: '/ (' ',9f12.4))
      write(6,101) ip
101   format(//' Cosines of angles between column vectors in',i3,
     *   ' dimensions')
      call maprnt(phi,m,m,mp,mp)
      return
      end
```

## BIBLIOGRAPHY

Note that a number of the references mentioned here were not referred to in the text directly, but were used as background reading as well as for cross-reference purposes. These are marked with an *.

Amari Shun-ichi (1982). Differential Geometry of curved exponential Families - Curvatures and Information Loss. *The Annals of Statistics*; Vol 10, no.2, pp 357-385.

Amari Shun-ichi (1985). *Differential-Geometrical Methods in Statistics.* Springer-Verlag. Berlin.

Anderson T.W. (1958). *Introduction to multivariate statistical Analysis.* New York: Wiley.

Anderson J.A. (1972). Separate sample logistic Discrimination. *Biometrika*; Vol 59, no. 1, 1972, pp 19-35.

Andrews D.F. and Pregibon D. (1978). Finding the outliers that matter. *Journal of the Royal Statistical Society, Series B*; Vol 40, pp 85-94.

Atkinson A.C. (1981). Two graphical Displays for outlying and influential Observations in Regression. *Biometrika*; Vol 68, pp 13-20.

Atkinson A.C. (1982a). Regression Diagnostics, Transformations and constructed

Variables. *Journal of the Royal Statistical Society,* Series B; Vol 32, pp 323-53.

Atkinson A.C. (1982b). Robust and diagnostic regression Analysis. *Communications in Statistics*; Series A, Vol 11, pp 2559-2571.

Atkinson A.C. (1985). *Plots, Transformations and Regression.* An Introduction to graphical Methods of diagnostic regression Analysis. Clarendon Press, Oxford.

Atkinson C. and Mitchell A.F. (1981). Rao's distance Measure. *Sankya*, Series A; Vol 43, pp 345-365.

Bard Y. (1974). *Nonlinear Parameter Estimation.* Academic Press.*

Barr G.D.I., Money A.H., Afflect-Graves J.F., Hart M.L. (1981). $L_p$-norm Estimation of the location Parameter of a symmetric Distribution. *South African Statistical Journal*; Vol 15, no 1, pp 85-96.*

Bates D.M. and Hamilton D.C. (1983). Calculation of intrinsic and parameter-effects Curvature for nonlinear regression Models. *Communications in Statistics - Simulations and Computations*; Vol 12, no 4, pp 469-477.*

Bates D.M. and Watts D.G. (1980). Relative Curvature Measures of Nonlinearity. *Journal of the Royal Statistical Society, Series B*; Vol 22, pp 41-88.

Bates D.M. and Watts D.G. (1981). Parameter Transformations for improved

approximate confidence Regions in non-linear Least Squares. *The Annals of Statistics*; Vol 9, no 6, pp 1152-1167.

Beale E.M.L. (1960). Confidence Regions in Non-linear Estimation. *Journal of the Royal Statistical Society, Series B*; Vol 42, pp 1-25.

Beckman R.J. and Cook R.D. (1983). Outliers.....s. *Technometrics*; Vol 25, no 2, pp 119-162.*

Belsley D.A. (1991). *Conditioning Diagnostics*. John Wiley & Sons, New York.

Belsley D.A., Kuh E. and Welsch R.E. (1980). *Regression Diagnostics: Identifying influential Data and Sources of Collinearity*. Wiley, New York.

Betts J.T. (1976). Solving the nonlinear least squares Problem: Application of a general Method. *Journal of optimization Theory and Applications*; Vol 18, no 4, pp 469-483.*

Box G.E.P. (1949). A general distribution Theory for a Class of likelihood Criteria. *Biometrika*; Vol 36, p 317.

Campbell N.A. (1978). The influence function as an aid in Outlier Detection in Discriminant Analysis. *Applied Statistics*; Vol 27, No. 3, 1978, pp 251–258.

Chalton D.O. (1990): *Contributions to Influence, Outliers and Bayesian Analysis in the multiple linear regression Model*. Ph.D. Thesis, University of Cape Town.

Chalton D.O. and Troskie C.G. (1992): Identification of outlying and influential Data with biased Estimation: A simulation Study. *Communications in Statistics – Simulation*; Vol 21, no 3, pp 607-626.

Chalton D.O. and Troskie C.G. (1993). On the Compatibility of sample and prior Information in the mixed regression model. In press.

Chambers John M. (1973). Fitting nonlinear Models: Numerical Techniques. *Biometrika*; Vol 60, no 1, pp 1-13.*

Clarke G.P.Y. (1980). Moments of the least squares Estimators in a nonlinear regression Model. *Journal of the Royal Statistical Society, Series B*; Vol 40, no 2, pp 227-237.*

Cook R.D. (1985). Residuals in nonlinear Regression. *Biometrika*; Vol 72, no 1, pp 23-29.*

Cook R.D. (1987). Parameter Plots in nonlinear Regression. *Biometrika*; Vol 74, no 4, pp 669-677.

Cook R.D. and Goldberg M.L. (1986). Curvatures for Parameter Subsets in Nonlinear Regression in *Computer Science and Statistics: The Interface*, Allen D.M. Editor. Elsevier Science Publishers, North-Holland.

Cook R.D. and Weisberg S. (1982). *Residuals and Influence in Regression.* Chapman and Hall, New York.

Cox D.R. and Hinkley D.V. (1974). *Theoretical Statistics.* Chapman and Hall, London.

Dawid A.P. (1975). Discussion to Efron's Paper. *Annals of Statistics*; Vol 3, pp 1231-1234.

De Jongh P.J., De Wet T. and Welsch A.H. (1988). Mallows-type bounded-influence-regression trimmed Means. *Journal of the American Statistical Association*; Vol 83, no 403, Theory and Methods, pp 805–810.

DiPillo P.J. (1979). Biased Discriminant Analysis: Evaluation of the optimum Probability of Misclassification. *Communications in Stat. - Theor. Meth*; Vol A 8(14), 1979, pp 1447–1457.

Donaldson J.R. and Schnabel R.B. (1986). Computational Experience with confidence Regions and confidence Intervals for nonlinear Least Squares in *Computer Science and Statistics: The Interface*, Allen D.M. Editor. Elsevier Science Publishers, North-Holland.

Draper N.R. and Smith H. (1981). *Applied regression Analysis*; 2nd ed. Wiley. New York

Du Toit S.H.C. and Gonin R. (1984). The Choice of an appropriate nonlinear Model for the Relationship between certain Variables in respiratory Physiology. *South African Statistical Journal*; Vol 18, pp 161-176.*

Eckart C. and Young G. (1936). The Approximation of one Matrix by another of

lower Rank. *Psychometrika*; Vol 1, pp 211–218.

Efron B. (1975). Defining the Curvature of a statistical Problem (with applications to second order efficiency). *The Annals of Statistics*; Vol 3, no 6, pp 1189-1242.

Fletcher R. (1980). *Practical methods of Optimization. Vol 1 - Unconstrained Optimization.* John Wiley. Chichester.*

Fisher R.A. (1925). Theory of statistical Estimation. *Proceedings of the Cambridge Philosophical Society*; Vol 122, pp 700-725.

Friedman Jerome H. (1987). Exploratory projection Pursuit. *Journal of the American Statistical Association*; Vol 82, pp 249–266.

Friedman Jerome H. (1989). Regularized discriminant Analysis. *Journal of the American Statistical Association*; Vol 84, no 405, Theory and Methods, pp 165–174.

Friedman Jerome H. (1991). Multivariate adaptive regression Splines. *The Annals of Statistics*; Vol 19, no 1, pp 1–141.

Forsythe G.E. and Moler C.B. (1967). *Computer Solution of Linear Algebraice Systems.* Prentice-Hall, Englewood Cliffs, N.J..

Fox J. (1984). *Linear statistical Models and related Methods.* Wiley & Sons, New York.

Gill P.E. and Murray W. (1974). Newton-type methods for unconstrained and linearly constrained Optimization. *Mathematical Programming*; Vol 7, pp 311-350.*

Gill P.E. and Murray W. (1978). Algorithm for the Solution of the nonlinear least square Problem. *SIAM Journal of Numerical Analysis*; Vol 15, no 5, pp 977-992.

Gnanadesikan R. and Kettering J. (editors) (1989). Discriminant analysis and clustering. *Statistical Science*; Vol 4, pp 34–69.

Gonin R. (1984). A Contribution to the Solving of nonlinear estimation Problems. *Research Report 1/84*; Ph.D.thesis.University of Cape Town.

Gonin R. and Money A.H. (1985(i)). Nonlinear $L_p$-norm Estimation. On the Choice of the Exponent, p, where the Errors are additive. *Communications in Statistics - Theory and Methods*; Vol 14, no 4, pp 827-840.

Gonin R. and Money A.H. (1985(ii)). Nonlinear $L_p$-norm Estimation. The asymptotic Distribution of the Exponent, p, as a Function of the sample Kurtosis. *Communications in Statistics - Theory and Methods*; Vol 14, no 4, pp 841-850.

Gonin R. and Money A.H. (1989). *Nonlinear $L_p$-norm Estimation.* Marcel Dekker, New York.

Good I.J. (1969). Some Applications of the singular Decomposition of a Matrix.

*Technometrics*; Vol 11, pp 823–831.

Green P.E. and Carroll J.D. (1976). ***Mathematical tools for applied Multivariate Analysis.*** New York: Academic Press.

Greenacre M.J. (1978). ***Some objective methods of graphical Display of a data Matrix.*** Doctoral thesis (Universeté Pierre et Marie Curie, Paris) published as a special report by the University of South Africa.

Greenacre M.J. (1980). ***Basic structure Display of a data Matrix.*** Research report 80/2, October 1980: University of South Africa.

Greenacre M.J. and Underhill L.G. (1981). Scaling a data Matrix in a low-dimensional Euclidian Space. In *Topics in Applied Multivariate Analysis*; Vol 2. Ed. D.M. Hawkins. Technical Report TWISK 193. Pretoria: Council for Scientific and Industrial Research, pp 239-336.

Greenacre M.J. and Underhill L.G. (1982). Scaling a data Matrix in a low-dimensional Euclidian Space. In *Topics in Applied Multivariate Analysis*; Vol 2. Ed. D.M. Hawkins. Cambridge: Cambridge University Press, pp 183-268.

Hastie T. and Tibshirani R. (1990). ***Generalized additive Models.*** Chapman and Hall, London.

Hastie Trevor, Tibshirani Robert and Buja Andreas (1992). Flexible discriminant Analysis. *Technical Report* - AT&T Bell Laboratories, New Jersey.

Hawkins D.M. (1980). *Identification of outliers.* Chapman and Hall, London.*

Hoaglin D.C. and Welsch R.E. (1978). The hat Matrix in Regression and ANOVA. *The American Statistician*; Vol 32, no 1, pp 17-22.

Jacobs M. (1983). *Linear regression Techniques for identifying influential Data and Applications in commercial data Analysis.* Unpublished Ph.D. Thesis at University of Cape Town.

John S. (1971). Some optimal multivariate Tests. *Biometrika*; Vol 58, pp 123–127.

Johnson R.A. and Wichern D.W. (1982). *Applied multivariate statistical Analysis.* Englewood Cliffs: Prentice-Hall.

Kass R.E. (1984). Canonical Parameterization and zero parameter effects curvature. *Journal of the Royal Statistical Society*; Series B, vol 46, pp 86-92.

Kendall M. and Stuart A. (1969). *The advanced Theory of Statistics*; Vol 1, London and Wycombe, Charles Griffin.

Kennedy W.J. and Gentle J.E. (1980). *Statistical Computing.* Dekker, New York.

Khorasani F. and Milliken G.A. (1982). Simultaneous confidence Bands for nonlinear regression Models. *Communications in Statistics - Theory and Methods*; Vol 11, no 11, pp 1241-1253.*

Koenker R.W. and Basset G.W. (1978). Regression Quantiles. *Econometrica*; Vol 46, pp 33–50.

Kotz and Johnson (Ed) (1983). *Encyclopedia of Statistical Sciences.* Vol 4, John Wiley & Sons, New York.

Kristof W. (1970). A Theorem on the trace of certain Matrix Products. *Journal of Mathematical Psychology*; Vol 7, pp 515–530.

Kshirsagar A.M. and Arseven E. (1975). A Note on the Equivalency of two discrimination Procedures. *The American Statistician*; Vol 29, No 1, pp 38–39.

Kuh E. and Welsch R.E. (1977). *Linear Regression Diagnostics*, Sloan School of Management Working Paper, pp 923-77. Massachusetts Institute of Technology, Cambridge, Massachusetts.

Lopuhaä Hendrik P. and Rousseeuw Peter J. (1991). Breakdown Points of affine equivariant Estimators of multivariate Location and covariance Matrices. *The Annals of Statistics*; Vol 19, no 1, pp 229–248.

MacNeill Ian B. and Jandhyala V.K. (1985). The residual Process for nonlinear Regression. *Journal of applied Probability*; Vol 22, pp 957-963.*

Mallows C.L. (1973). *Influence Functions*; unpublished paper presented at a conference on robust regression held at Cambridge Mass.

Mallows C.L. (1975). *On some Topics in Robustness*; unpublished memorandum,

Bell Telephone Laboratories, Murray Hill, New Jersey.

Marquardt D.W. (1963). An Algorithm for least squares Estimation of nonlinear Parameters. *Journal of the Society for industrial and applied Mathematics - (Numerical Analysis)*; Vol 11, pp 431-441.

Marquardt D.W. (1970). Generalized Inverses, Ridge Regression, biased linear Estimation, and nonlinear Estimation. *Technometrics*; Vol 12, no 3, pp 591-611.

Marshall A.W. and Olkin I. (1968). Norms and Inequalities for condition numbers, III. *Technical Report* No 53. Stanford University.

Militkŷ Jiřı̌ and Jaroslov Cãp (1985). Detection of influential Points in general nonlinear Procedures. *Problems, System Analysis and Simulation*; Vol 27, pp 365-368.*

Milliken G.A. (1978). A Procedure to test Hypoteses for nonlinear Models. *Communications in Statistics - Theory and Methods - A*; Vol 7, no 1, pp 65-79.*

Money A.H., Affleck-Graves J.F., Hart M.L. and Barr G.D.I. (1982). The linear regression Model: $L_p$-norm Estimation and the Choice of p. *Communications in Statistics-Simulation and Computations*; Vol 11, pp 89-109.

Moolgavkar Sureh H., Lustbader Edward D. and Venzon David J. (1984). A geometric Approach to nonlinear regression Diagnostics with Application to matched case-controlled Studies. *The Annals of Statistics*; Vol 12, no 3, pp

816-826.*

Morrison D.F. (1976). *Multivariate Statistical methods*, 2nd Edition. New York: McGraw-Hill.

Nazareth L. (1980). Some recent Approaches to solving large residual nonlinear least squares Problems. *SIAM Review*; Vol 22, no 1, January 1980.

Newhouse J.P. and Oman S.D. (1971): An Evaluation of Ridge Estimators. *Rand Report* No R-716-PR (April, 1971), pp 1-28.

Olkin Ingram (1992). A matrix Formulation of how deviant an Observation can be. *The American Statistician*, Vol 46, no 3, pp 205-209.

Osborne M.R. and Watson G.A. (1971). On an Algorithm for discrete non-linear $L_1$-approximation. *Computational Journal*; Vol 14, pp 184-188.

Pillai K.C.S. (1957). *Concise Tables for Statisticians*. Bockman, Inc. Manilla.

Pregibon D. (1980). Goodness of link tests for generalized linear Models. *Applied Statistics*; Vol 29, pp 15-24.

Pregibon D. (1981). Logistic regression Diagnostics. *The Annals of Statistics*; Vol 9, no 4, pp 705-724.

Press S.J. (1972). *Applied multivariate Analysis*. Holt, Rinehart and Winston, Inc., New York.

Rao C.R. (1945): Information and Accuracy attainable in the Estimation of Statistical Parameters. *Bulletin of the Calcutta Mathematical Society*; Vol 37, pp 81–91.

Rao C.R. (1963). Criteria of Estimation in large Samples. *Sankhyā*; Vol 25, pp 198-206.

Rao C.R. (1973). *Linear statistical Inference and its Applications* (2nd ed.) John Wiley, New York.

Ratkowsky D.A. (1983). *Non-linear regression Modelling.* Marcel Dekker, New York.

Reed J. (1975). Discussion to Efron's Paper. *Annals of Statistics*; Vol 3, pp 1234-1238.

Rosner B. (1983). Percentage Points for a generalized ESD many outlier Procedure. *Technometrics*; Vol 25, no 2, pp 165-127.*

Ross W. (1984). *Measuring Influence in nonlinear Regression.* Unpublished Thesis. Queen's University, Kingston Canada.

Rousseeuw Peter J. and Van Zomeren Bert C. (1990). Unmasking multivariate Outliers and leverage Points. *Journal of the American Statistical Association*; Vol 85, no 411, Theory and Methods, pp 633–651.

Schwetlik H. and Tiller V. (1985). Numerical Methods for estimating Parameters

in nonlinear Models with errors in the Variables. *Technometrics*; Vol 27, no 1, pp 17-24.*

Skovgaard L.T. (1984). A Riemannian Geometry of the multivariate normal model. *Scandinavian Journal of Statistics*; Vol 8, pp 227-236.

Sposito V.A. (1982). On unbiased $L_p$ regression Estimators. *Journal of the American Statistical Association*; Vol 77, pp 652-654.

Thiart C., Dunne T.T., Troskie C.G. and Chalton D.O. (1991). A simulation Study of biased Estimators against the ordinary Least Square Estimator. *Technical Report*; University of Cape Town.

Thomas T.Y. (1961). *Concepts from Tensor Analysis and Differential Geometry.* Academic Press, New York.*

Troskie C.G. (1977). Multicollinearity, Ridge Regression and Principal Components. *Unpublished Report.* University of Cape Town.

Troskie C.G. and Conradie W.J. (1986). The Distribution of the ratios of Characteristic Roots (Condition Numbers) and their Application in Principal Component or Ridge Regression. *Linear Algebra and its Applications*; Vol 82, pp 255–279.

Underhill L.G. (1988). *Private Communication.* University of Cape Town.

Van Deventer P.J.U. (1985). *The Applicability of Discriminant Analysis*

*Techniques on the multivariate normal and non-normal data Types in marketing Research.* Unpublished M.Sc. Thesis, University of Cape Town.

Vetterling W.T., Teukolsky S.A., Press W.H. and Flannery B.P. (1985). *Numerical Recipes: The Art of Scientific Computing.* Cambridge University Press.

Vinod H.D. (1978). A Survey of Ridge Regression and related Techniques for Improvement over ordinary least Squares. *Review Econ. and Statist.*; LX, pp 121–131.